DOCUMENT RESUME

ED 299 320 TM 012 325

AUTHOR Tatsuoka, Kikumi K.
TITLE Toward an Integration of Item-Response Theory and Cognitive Error Diagnoses.
SPONS AGENCY Office of Naval Research, Arlington, Va. Personnel and Training Research Programs Office.
PUB DATE [87]
CONTRACT N00014-82-K-0604
NOTE 70p.; This paper is a chapter in the book "Diagnostic Monitoring of Skill and Knowledge Acquisition," Frederiksen, Glaser, Lesgold, and Shafto (Editors); and is based on a paper presented at the Educational Testing Service Conference (Princeton, NJ, July 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Bayesian Statistics; *Cognitive Processes; Elementary Education; Equations (Mathematics); *Latent Trait Theory; Misconceptions; *Problem Solving; *Psychometrics; Test Construction
IDENTIFIERS *Error Analysis (Statistics); Graph Theory; *Rule Space Model

ABSTRACT

The Rule Space Model, a cognitive error diagnostic methodology, is discussed, and the philosophy behind it is related to the question of what really determines item response curves. The Rule Space approach integrates Item Response Theory and the algebraic theory of databases. An application of Graph Theory is introduced as a way to acquire a list of bugs to be diagnosed--sources of misconceptions, locations of incomplete knowledge, lack of knowledge, or combinations of task attributes relevant to solutions of problem. Rules in the bug information bank are translated into multivariate normal distributions (bug distributions) in the rule space, and their behaviors are discussed in the context of statistical profiles of the bug distributions. The rule space is shown to be useful in diagnosing different strategies adopted by students for solving the same problems, as well as in diagnosing individual cognitive errors. Diagnoses are carried out using Bayes' decision rule for minimum error after computing Mahalanobis distances of the students' point to the centroids of ellipses representing the bug distributions. Ten figures and five tables are provided. A list of 89 references concludes the document. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED299320

TOWARD AN INTEGRATION OF ITEM-RESPONSE THEORY AND COGNITIVE ERROR DIAGNOSES

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

KIKUMI TATSUOKA

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Kikumi K. Tatsuoka

University of Illinois at Urbana-Champaign

Computer-based Education Research Laboratory

A chapter in

Diagnostic Monitoring of Skill and Knowledge Acquisition

Frederiksen, Glaser, Lesgold and Shafto (Eds.)

based on a paper given at the Educational Testing Service Conference

Princeton, NJ, July, 1986

TM012325

Acknowledgement

This research was sponsored by the Personnel and Training Research Programs, Psychological Sciences Division, Office of Naval Research, under Contract No. N00014-82-K-0604, Contract Authority Identification Number NR 150-495.

The author wishes to acknowledge Robert Baillie for computer programing, Betty Dowd and Kentaro Yamamoto for collecting data and miscellaneous assistance, Jocelyn Smith and Rose Nygaard for typing this manuscript and drawing the charts, and Maurice Tatsuoka for his critical review and editorial help.

The author is especially grateful to Bob Linn, Robert Mislevy, Bill Stout and Maurice Tatsuoka for valuable discussions and comments.

The data used in this report was obtained from a study supported by the National Institute of Education.

Some of the analyses presented in this report were performed on PLATO^{R} system. The PLATO^{R} system is a development of the University of Illinois, and PLATO^{R} is a service mark of Control Data Corporation.

The NovaNET^{SM} system is a development of the University of Illinois and NovaNET^{SM} is a service mark of University Communications, Inc..

Abstract

This study introduces an outline of the cognitive error diagnostic methodology (Rule space model) developed over the past several years by Tatsuoka and her associates. Objectives of the theory of cognitive error diagnosis are discussed first, and then pros and cons of latent trait and latent class models are discussed in terms of four dimensions of the goals to be attained in modern educational measurement theory. This paper is intended to clarify the philosophy behind the rule space model and tries to relate it to one of the fundamental issues of educational measurement--what really determines the item response curves, a question which is somewhat similar to Caroll's quest for "what ability is." A new technique that is an application of Graph Theory is introduced to acquire a list of the bugs to be diagnosed--sources of misconceptions, locations of incomplete knowledge or lack of knowledge in a given knowledge structure, or various combinations of task attributes relevant to the solution of problems. The rules listed in the bug information bank are translated into multivariate normal distributions (bug distributions) in the rule space, and their behaviors are discussed in the context of statistical properties of the bug distributions. Finally, it is shown that the rule space is useful in diagnosing different strategies adopted by students for solving the same problems, together with diagnosing individual cognitive errors. Diagnoses are carried out by using Bayes' decision rule for minimum error after computing the Mahalanobis distances of the student's point to the centroids of the ellipses representing the bug distributions.

Introduction

Finding the sources of misconceptions possessed by students is a difficult task because it is impossible to see what is happening in their heads. The only directly observable outcomes are the students' responses to the test items. Studying their think-aloud protocols is one method for discovering how students solve or think through problems. Several computer programs that are capable of diagnosing students' misconceptions have been developed in the past decade (Brown & Burton, 1978; Marshall, 1980; VanLehn, 1983; Tatsuoka, Baillie, & Yamamoto, 1982; Sleeman, 1984; Ohlsson & Langley, 1985). The common ground for these cognitive diagnostic systems is that they infer the unobservable cognitive processes from logical interrelationships among cognitive tasks, subtasks and goals involved in problems representing the domain of interest. It is important that we be able to retrieve invisible things from the "black box" and put them into a useful form so that valuable information can be obtained for improving educational quality.

Efforts to incorporate the findings of cognitive or educational theories into the construction of test items is not new. Bloom and his associates (1956), for example, created a taxonomy of educational objectives for the cognitive domain, and it has become very popular among educational practitioners in the past decades. However, some people have had considerable difficulty in classifying test items according to Bloom's taxonomy. Statistical properties of the taxonomy have been investigated (Madaus, Woods & Nuttall, 1973;

Miller, Snowman, & O'Hara, 1979), and the results suggested only two cognitive factors--fluid and crystallized intelligence (Cattell, 1971).

More recent advances in cognitive theory provide new insights into human thinking and learning processes. As a result, it had become apparent that the need for a new kind of test has arisen--tests that may be used as an integral part of instruction. In order to be of greater use, the new assessments should promote learning and the improvement of instruction.

Linn (1985) pointed out that "there are a number of barriers that will need to be overcome if the envisioned improvements in testing are to be even partially realized." He describes three barriers: a lack of technical or methodological theories that are appropriate to handle dynamic aspects of modern learning theory, an economic barrier, and an ideological barrier. The ideological barrier has to do with the researchers and practitioners who in dealing with current theories of educational and psychological statistics perceive "noise" or "response errors" as incidental factors, and fail to seek reasons for the errors. Aberrant responses may be seen as errors even though they coincide with the responses resulting from the perfect application of wrong rules (Birenbaum and Tatsuoka, 1982). Linn (1985) says that "it is important to understand the obstacles to change, whatever their nature, in order to overcome them."

In this chapter, a new methodology that is capable of diagnosing cognitive errors and analyzing different methods for solving problems

will be introduced and illustrated with fraction subtraction problems. This new approach, called "Rule Space," integrates Item Response Theory (IRT) and the algebraic theory of databases (Lee, 1983). The rule space model is general enough to apply to any domain where classification and selection of students and diagnoses of misconceptions are done by Bayes' decision rules for minimum errors (Fukunagawa, 1972), or where any other equivalent decision rules are applied to a set of systematic errors determined prior to the analyses. The first section of the chapter discusses important objectives and criteria for the construction of cognitive diagnostic tests must follow. Then the rule space model will be introduced and related to a distribution theory, to the construction of a "bug" library and a rule space, and finally to an operational classification scheme.

On the Construction of Items for Cognitive Diagnostic Testing

Special care has to be taken in selection of the items for cognitive diagnostic testing. For example, one of the popular misconceptions in fraction subtraction problems is held by about 3% of seventh or eighth graders. When it is necessary to increase the numerator of the first fraction of a subtraction problem, a student with the misconception reduces the whole number part by one and adds 10 to the numerator of the fraction (Shaw, 1984) (presumably by analogy with borrowing in whole number subtraction). But if the denominator of the first fraction happens to be 10 the procedure produces the correct answer.* Therefore, in order to detect the

*$\left[2\frac{5}{10} - \frac{6}{10} = 1\frac{15}{10} - \frac{6}{10} = 1\frac{9}{10},\ 2\frac{1}{5} - \frac{2}{5} = 1\frac{11}{5} - \frac{2}{5} = 1\frac{9}{5} \right]$

misconception, at least one item must have a denominator not equal to 10 in the first fraction, while another item must have the denominator of 10. Then the response pattern (1 for the correct answer and 0 for wrong answers) yielded by the right rule is (1,1), but the rule adding 10 to the numerator produces the response pattern of either (1,0) or (0,1).

Traditionally, item construction was based on the concept of content validity, that is, "how well the content of the test item covers the class of situations or subject matter about which conclusions are to be drawn" (APA, 1974). However, Angoff (1986), Messick (1980), and Cronbach (1975) came to the same conclusion offered by Loevinger in 1957: "Content validity is essentially ad hoc and does not have scientific value." The recent view of content validity is generally replaced by construct validity.

To establish construct validity it is necessary to "examine the psychological trait, or construct, presumed to be measured by the test and investigate relationships between the data from the test and the theory underlying the construct" (Messick, 1984). Further, Angoff (1986) states that "Construct validation is a process, not a procedure; and it requires many lines of evidence, not all of them quantitative." According to his view, a validation study may include logical task analyses such as that done by Klein, Birenbaum, Standiford and Tatsuoka (1981) for predicting bugs or erroneous rules of operation, and the construction of test items by which all the predicted bugs will be covered. Protocol studies have also become an important component of the validation process.

Glaser (1985) suggested a new direction for educational measurement in achievement test. He writes, "Test items can be comprised of two elements--information that needs to be known and information about the conditions under which use of this knowledge is appropriate." There are obviously various stages of competence in students' knowledge, including cognitive skills. And as Glaser implies, it is important to discover what knowledge structures the students have. Greeno (1980) pointed out that the acquisition of declarative and procedural knowledge is usually an object of instruction, but strategic knowledge that enables one to set goals and subgoals and to form plans for attaining goals is not explicitly taught. It is often left to individuals to acquire it by induction. Some item types require the students to decide which solution path should be taken, beginning with what should be done first to reach the final answer. Many erroneous rules discovered in the studies of signed-number and fraction addition and subtraction problems indicate that the students do not even recognize that different item types require different solution paths and different sequences of subtasks. They did not have the slightest idea of why the strategic skill described by Greeno were necessary.

The new test design must be capable of detecting the different knowledge structures possessed by individuals. Each different structure requires its own unique strategic skills. For example, there are two contrasting methods for solving mixed number fraction subtraction problems. One is to separate whole number parts from fraction parts (Method B), and the other is to convert mixed numbers

to improper fractions first and then subtract two numbers (Method A). If a student using Method A has excellent computational skills in the arithmetic of whole numbers, then he/she does not have to learn how to borrow one from the whole number part. So Method A always gives the correct answers for subtraction of mixed numbers. The student using Method A can get high scores without really understanding numbers and the number system.

Modern cognitive theory concludes that one of the important stages in learning processes is "theory change" (Glaser, 1985). When learning takes place, students evaluate their hypotheses and then modify as necessary their current theories on the basis of the new information. New methods of educational measurement must take into account such volatile theory changes by an individual and be able to capture the traces of performance changes in detail.

The goals to be attained in modern measurement theory are not easy. As Linn states, the technical barrier is high and the traditional theories of educational measurement and testing have only limited power, or are simply inapplicable to the new goals. IRT is not an exception if we conceptualize IRT at the level of individual items.

Thus the usual definition of construct validity has to be expanded and traditional test theory needs to be reconceptualized in order to explain the dynamic aspects of learning and to express knowledge structures in terms of relational databases. None of the traditional test theories can handle theory changes or assess knowledge structures.

Glaser (1985) categorizes the main objectives of assessing new achievement measures into four parts: The first is to diagnose the principles of performance, the second is to assess theory changes, the third is to evaluate the structure or representation of problems, and the fourth is to assess the automaticity of performing skills. The automaticity is important to reduce demands on attention. Carrying out single component processes may be easy, but it may not be easy to work on several components together unless one or more have become automatic. The importance of such orchestration of several component tasks seems often to be neglected by textbooks and in classroom teaching.

The four objectives listed above simply need new achievement measures that are descriptive and qualitative. Psychometrics, on the other hand, is concerned mainly with quantitative theories of educational and psychological measurement. Test theory has supported the development of standardized testing throughout a long history of contributions to American education, and the development of IRT has led many improvements in such area as item analysis, test design, test equating, and procedures for detecting item bias. Test theory should also be useful in supporting the development of the new achievement measures.

The basic concept underlying IRT is latent traits. As stated by Lord and Novick (1968, p. 358), "A theory of latent traits supposes that an individual's behavior level can be accounted for, to a substantial degree, by defining certain human characteristics called traits, quantitatively estimating the individual's standing on each

of these traits, and then using the numerical values obtained to predict or explain performance in relevant situations." Messick (1984) projected his view of new achievement assessment as a combination of trait theory and the descriptive theory used in differential psychology. However, there is no guarantee that traits exist in any physical or physiological sense. "It is sufficient that a person behaves as if those amounts substantially determined his behavior," according to Lord and Novick (1968, p. 358). Convincing interpretation of trait variables have never been given in the past.

Returning to the definition of construct validity given by Angoff (1986), let us examine the psychological traits or constructs as viewed by traditional psychology. Cattell describes ability as organized complexes of transferable concepts and skills. Guilford (1967) describes it as information-processing skills. Sternberg (1977) treats abilities as constellations of information-processing components. Snow & Lohman (1984) conceive abilities as structures of assembly and control processes as well as performance processes.

On the other hand, psychometricians have developed probabilistic models to measure invisible constructs. Two general classes of the models have been proposed: continuum and state models. Two conflicting views are at the root of the models: For continuum models, trait acquisition is assumed to be continuous in nature, while state models take the position of all-or-none, discrete processes. Continuum models would appear to be more relevent to supporting the learning tests of the future.

As for the continuous models, several pioneers have developed

various types of item response theory models (Lord, 1953; Lord and Novick, 1968; Birnbaum, 1968; Rasch, 1960). Since then, several extentions or modifications of the original IRT models have been developed (Samejima, 1969, 1983; Bock, 1972; Fischer, 1973; Embretsen, 1984; Mokkon & Lewis, 1982; Reckase, 1985; Mislevy, 1983). The IRT models express a trait by the variable θ and express item characteristics by continuous probability functions. That is, item scores $x_j(j=1,2...n)$ are related to a trait θ by a function that gives the probability of each possible score on an item for a randomly selected examinee of given ability. These functions are response curves (formerly called item characteristic curves by Lord & Novick, 1968). Although the models fit the datasets of various standardized tests, the clear meaning of θ remains undetermined through data analysis. Spada and McGaw (1985) investigated cases for which the one-parameter logistic model and Fischer's linear logistic test model (LLTM) are applicable. They concluded that "the simple Rasch model is applicable only if global learning of item-specific learning occurs, with constant gains for all persons. Person-specific learning falsifies the model. The same is true for application of LLTM that decompose item-specific learning into changes in the difficulties of elementary cognitive operations" (p. 189).

A discrete learning model pioneered by Lazarsfeld (1956) and Lazarsfeld & Henry (1968) also has branched out to many modified, extended models (Macready, 1982; Alvord and Macready, 1985; Muthén, 1985; Paulson, 1985; Yamamoto, 1987). Paulson, in particular,

extended the latent structure model to apply to the detection of rules of operation in signed-number addition problems. Each rule was treated as a discrete state in his model. Some basic assumptions in the state models are too restrictive and unrealistic to incorporate into the modern learning theory. They assume a priori how many latent classes or states the model has. Then, every subject must belong to exactly one of the finite sets of classes which are mutually exclusive and together exhaustive. Thus the explanation of theory changes by state models is conceptually very difficult. The restriction in the number of states to be included in modeling before the parameters are estimated critically limits the flexibility of this approach. Although recent advances in methods of estimating parameters (Bock & Aitken, 1981; Bock & Mislevy, 1982) are significant, it is still an expensive task with respect to computer time, and it requires a dataset of astronomical size in order to obtain accurate estimates of many parameters.

Tatsuoka and her associates (Tatsuoka, 1983, 1985, 1986; Tatsuoka & Tatsuoka, 1983, in press) have developed a new probabilistic model which has taken advantage of both the continuous and discrete models. They named it the "Rule Space" model.

The Rule Space Model

1. Philosophy behind the rule space model

Whatever the invisible traits or constructs stand for, the statistical meaning of the estimated latent variable is virtually equivalent to the proficiency levels of the performances on the test items, because the total score or weighted total score is a

sufficient statistic for estimating the true θ by the maximum likelihood method in the one- and two-parameter logistic models (two of the basic IRT models) and nearly so in the three-parameter logistic and other related IRT models.

Suppose item j is scored 1 or 0; then x_j is a random variable related to θ by a probability function as follows:

$$(1) \quad P_j(\theta) = \text{Prob}(x_j = 1|\theta) = 1 - Q_j(\theta) = 1 - \text{Prob}(x_j = 0|\theta).$$

The one-parameter logistic function defines a basic type of model called the Rasch Model. Equation (2) gives the two-parameter logistic function, where a_j is item discrimination power and b_j is item difficulty:

$$(2) \qquad P_j(\theta_i) = 1/\left[1 + \exp(-1.7a_j(\theta_i - b_j))\right].$$

If a_j is set equal to 1, then it becomes the Rasch model.

Tatsuoka and Linn (1983) discussed the relationship between item response function $P_j(\theta_i)$ and person response function $P_i(b_j)$. The person response function (or curve) is defined by the same equation (2), but $P_j(\theta_i)$ is a function of a continuous variable θ for fixed b_j, a_j (and other parameters such as c, if present), while $P_i(b_j)$ is a function of variable b (assuming there are infinitely many items) for a fixed level of θ_i and fixed a. In particular, the one-parameter logistic function is a symmetric function with respect to θ and b.

The person response function is the probability function of person i with $\theta_i = \theta$ getting the correct answer for an item with difficulty b_j. Since Mosier explored person response curves in 1941,

several researchers have found them very useful for explaining the relation between ability θ and item difficulty b (Trabin & Weiss, 1979; Carroll, 1985). By using the one-parameter logistic model, Carroll (1987) explored the relation between both the curves with several ability tests in order to obtain an answer to the question, "what is an 'ability'?" (p. 1). His assumption is that "The existence of an ability can be demonstrated when it can be shown that for any individual there is a systematic, monotonic, and close relation between the individual's probability of correct or satisfactory performance and the difficulties of a series of tasks, and when there are variations over individuals in the parameters of this relation." (p. 22). One of his conclusions is that the ability is defined in terms of the attribute(s) of the tasks that cause differences in task difficulty. Carroll's conclusion is that the ability is defined in terms of the attribute(s) of the tasks that cause differences in task difficulty. This conclusion is applicable to the situation such that the space of the difficulties resulting from various combinations of attributes involved in the tasks is unidimensional. Then his conclusion is mathematically sound because of the symmetric relation of the Rasch model.

As long as we interpret θ as the latent ability or construct which influences the performances on the tests, we may face the philosophical dilemmas of IRT models, such as the dimensionality of θ or b, or the impossibility of explaining gain scores or Glaser's theory changes. Since a composite of several factors (or abilities) influences the performance of an item, and each item in the test is

likely to require a different composite or possibly a different set of abilities, the psychological meaning of θ is very complicated and difficult to interpret (Stout, in press).

IRT models are formulated by utilizing a response pattern or binary vector with n elements. There are two distinct independent pieces of information in a response pattern. One is quantitative, telling us how many items are correctly answered (total score), and the other is qualitative, telling us which items are correctly answered and which are missed. Since IRT assumes local independence (Lord & Novick, 1968), the likelihood function is expressed in terms of both pieces of information. The item and person parameters can be estimated from the likelihood function by applying the maximum likelihood procedure.

However, Tatsuoka (1987) investigated what really determines the item response curves, a task somewhat similar to Caroll's quest regarding "what ability is." Starting from a painstaking logical task analysis, she constructed sets of items such that each item involves a unique combination of cognitive subtasks. The study showed that the underlying cognitive processes for solving an item determine the slope and location parameters of an IRT curve. If this is true for any domain, then it is important to expand our view from a localized, narrow interest in IRT curves to a broader global concept. After IRT models were introduced, refinement and improvement in psychometric techniques have generally been limited to dealing with individual items separately.

A new model has to be able to achieve the measurement objectives

of modern learning theory. In order to achieve this demanding goal, it is helpful to see item response curves as a whole, as a set, and to investigate algebraic and topological properties of this function space. Many IRT models have been proposed recently, and their probability functions provide a finer or more accurate measure of θ than simple basic IRT models. They can also be used to formulate the rule space model; so it is not necessary to be restricted only to one- and two-parameter logistic models. Such an expansion of the perspectives of test theories to functional analysis leads to some important and useful conclusions. For example, Tatsuoka (1975) reformulated classical test theory in a vector, Hilbert space and proved the existence of the true score.

The theory of functional analysis treats a function as a "point" and utilizes the methods of algebra and topology in a set of functions. Ramsay (1982) outlined the concept of functional analysis as an extension of classical statistical techniques and explained least squares, principal components, and canonical correlation analyses in the context of functional analysis. He stated that "the data must be viewed as an element in a space of possible functions taking a domain space into a range space" (p. 352). He concludes with the remark that "functional analysis already has revolutionized numerical analysis so that any issue of a major journal [in that discipline] now has a number of papers using this technology. I claim that this is about to happen to statistics" (p. 394). The rule space model is an application of functional analysis using a projection operator.

The third objective listed by Glaser was to assess the structure of knowledge possessed by an individual. This dimension requires a leap to a new world for the field of educational measurement, as it is apparent that the algebraic theory of relational databases is needed to achieve this objective. If a set of items is carefully constructed, then various relationships among the items should reflect the relationships among cognitive subtasks underlying each item. By examining bugs and sources of misconceptions, one can see which subtasks caused scores of ones or zeros on the items. For example, Tatsuoka (1984c) represented 27 erroneous rules of operation found in signed-number addition problems as ordered pairs of (1) the number of cognitive steps taken correctly, and (2) the value of the Norm Conformity Index, one of several indices that measure appropriateness of response patterns (Tatsuoka & Tatsuoka, 1982). The latter assesses the degree of conformity of the sequence of cognitive steps followed by each bug to the expert's procedural steps. That is, the second value measures how early or how late an erroneous rule causes departure from the correct steps in the procedural network. The result of the study indicated that an early derailment has more serious consequences than one at a later stage. Cluster analysis separated bugs caused by early and frequent derailments from those due to later and less frequent derailments. She named the former "seriously ill-composed rules." Later studies (Birenbaum & Tatsuoka, 1986a; Shaw, 1986) indicated that wrong rules that are not seriously ill-composed can be remediated by giving correct answers, or even by the feedback of OK or NO to each reponse to the item.

The Norm Conformity index characterizes the quality of information of a response pattern and expresses its characteristics quantitatively. The rule space uses a similar index (called IRT-based caution index) defined in the context of the IRT curves (Tatsuoka, 1984b). However, in order to explain what the rule space model is, we have to clarify the terms "bug" and "erroneous rules of operation," which have been used without specific definitions so far.

2. What are Rules?

A systematic error over the test items can result from a combination of one or more bugs or erroneous rules. If a student applies his/her erroneous rule with perfect consistency to the items in the test, then his/her responses to the test will be perfectly matched with the responses generated by a computer program. We call systematic errors erroneous rules.

Correct rules will, by definition, produce the right answer to all the items, but sometimes wrong rules may produce the right answer to some subset of the test items. Moreover, some rules are combinations of the right rule and wrong rules, and others are combinations of two or more different wrong rules. We consider them as new rules as long as they are consistently applied to all the items. If we construct the test items carefully so that the important, predicted common errors can be expressed by unique item response patterns of ones and zeros, or component response patterns specified in a task analysis, then the rules can be distinguished by response patterns. In other words, we can assume that rules are expressed by binary response vectors.

The rule space model was developed to solve a specific classification problem in which the entities to be classified are the rules in some well-defined domain such as arithmetic, algebra or science. As was mentioned earlier, two kinds of information are included in a response pattern. They are the quantitative information of the total number of ones and the qualitative information on which items had ones. The former is represented by θ, the latter, by ζ, an index expressing atypicality (or inappropriateness) of the response patterns of a given group (Tatsuoka, 1984b). Any rule can be expressed by an ordered pair of θ and ζ or, equivalently, θ and the numerator of ζ, denoted by $f(\underset{\sim}{x})$ (Tatsuoka, 1985). We assume that, at least at the very beginning, the rules are determined a priori through a logical task analysis. A later section explains this in detail.

3. Bug Distribution

The distribution of observations plays an important role in statistical theories. When we deal with students, random errors or slips due to careless errors or uncertainty always affect the outcomes of performance on a test. A slip on item j occurs when a student does not use his/her rule exactly on item j, and as a result, the response to item j does not match the response yielded by the rule. The response which deviated from the rule-based response may, of course, be the correct response. Even if a student possesses some systematic error, it is very rare to have the response pattern perfectly matched with the patterns theoretically generated by its algorithm (VanLehn, 1983; Tatsuoka, 1984a). Some systematic errors

may produce more slips, while others have a small number of slips. Also, some items may be prone to produce more slips than other items. It is very important that we be able to predict the probability of having slips on each item for each systematic error (or rule). Knowing the theoretical distribution of observed slips of a rule enable us to see and predict statistical properties of observed responses yielded by the rule.

Tatsuoka and Tatsuoka (1987) derived the theoretical distribution of observed slips and named it the "Bug Distribution." First, the probability of having a "slip" on item j (j=1,2,...,n), is denoted by p_j for item j, (j=1,2,...,n) and it is assumed that slips occur independently across items. The bug distribution of a rule R follows a compound binomial distribution with different slip probabilities for the items.

$$(3)\quad \text{Prob (having up to s slips}|R) = \left\{ \sum_{m=0}^{s} \left\{ \sum_{\Sigma u_j = m} \prod_{j=1}^{n} p_j^{u_j} (1 - p_j)^{1-u_j} \right\} \right\}$$

where u_j is a random variable such that $u_j = 1$ if a slip occurs on item j and $u_j = 0$ if not. The expectation and variance of the bug distribution of Rule R whose corresponding binary vector is $\underset{\sim}{x}_R$--in which we assume, without loss of generality, that the first r elements are ones and the remainder are zeros (i.e., $x_{Rj}=1$ for j=1,...,r and $x_{Rj}= 0$ for j=r+1,...,n)-- will be given by

$$(4) \qquad \mu_R = \sum_{j=1}^{r} p_j + \sum_{j=r+1}^{n} q_j$$

$$(5) \qquad \sigma_R^2 = \sum_{j=1}^{n} p_j q_j$$

where $q_j = 1 - p_j$.

The bug distribution of rule R is a function defined on the slip random variable u_j, and it relates algebraically to the conditional probability that a subject in the state of possessing rule R will respond correctly to item j (Tatsuoka & Tatsuoka, 1987b).

4. A Mapping Function of Response Patterns $\underset{\sim}{x}$.

The rule-space model begins by mapping all possible binary response patterns, $\underset{\sim}{x}$'s, into a set of ordered pairs $\{(\theta, f(\underset{\sim}{x}))\}$. The mapping function $f(\underset{\sim}{x})$ is the inner product of two residual vectors, $\underset{\sim}{P}(\theta) - \underset{\sim}{x}$ and $\underset{\sim}{P}(\theta) - \underset{\sim}{T}(\theta)$, where $P_j(\theta)$ $j=1,\ldots n$ are the logistic functions and $\underset{\sim}{P}(\theta) = \left[P_1(\theta),\ldots P_n(\theta)\right]$ and $\underset{\sim}{T}(\theta) = \left[T(\theta),\ldots T(\theta)\right]$, $T(\theta)$ being the average of $P_j(\theta)$, $j=1\ldots,n$. Since $f(\underset{\sim}{x})$ is a linear function, the bug distribution of Rule R mapped into the rule space will have the centroid $(\theta_R, f(\underset{\sim}{x}_R))$, where $f(\underset{\sim}{x}_R)$ is given by Equation (6), and the variance-covariance matrix of Equation (7) (Tatsuoka, 1985).

$$(6) \quad f(x_R) = -\sum_{j=1}^{r} Q_j(\theta_R)\left[P_j(\theta_R) - T(\theta_R)\right] + \sum_{j=r+1}^{n} P_j(\theta_R)\left[P_j(\theta_R) - T(\theta_R)\right]$$

$$(7) \qquad \underset{\sim}{\Sigma}_R = \begin{bmatrix} 1/I(\theta_R) & 0 \\ 0 & \sum_{j=1}^{n} P_j(\theta_R)Q_j(\theta_R)(P(\theta_R) - T(\theta_R)) \end{bmatrix}$$

where θ_R is the θ-value for Rule R. $Q_j(\theta_R)$ is $1 - P_j(\theta_R)$ and $I(\theta_R)$ is the information function of the test at θ_R.

On the other hand, the slip variable u_j is related to the item-score variable x_j on item j as follows:

(8) $\begin{cases} u_j = 1 \text{ if a slip occurs and hence } x_j \neq x_{Rj} \\ u_j = 0 \text{ if a slip does not occur and hence } x_j = x_{Rj}. \end{cases}$

or, equivalently $u_j = |x_{Rj} - x_j|$ by Tatsuoka & Tatsuoka (1987b). The conditional probability of having a slip on item j for a given rule R is as follows:

(9) Prob $(u_j = 1 \ |R) =$ Prob $(x_j \neq x_{Rj}|R)$.

Tatsuoka (1985) showed that the function $f(\underset{\sim}{x})$ maps binary response patterns $\{\underset{\sim}{x}\}$ into a set of ordered pairs $\{\theta, f(\underset{\sim}{x})\}$ one-to-one in the one-parameter logistic model, and also empirically confirmed with several sets of Monte Carlo data that this relation would be true for the two-parameter logistic model as well, provided the number of items is large and the a_j are spread out. Therefore, the relationship in (9) can be rewritten, within the context of IRT models, in Equation (10),

(10) Prob $(x_j \neq x_{Rj} \mid R) =$ Prob $(x_j \neq x_{Rj} \mid \theta_R, f(\underset{\sim}{x}_R))$.

The relation (10) is true for any IRT model as long as the mapping of response patterns $\{\underset{\sim}{x}\}$ into ordered pairs of ability level θ and the image of $\underset{\sim}{x}$ is one-to-one. The mapping function that is used to formulate the rule space can be other person fit statistics such as those of Drasgow, Levine & McLaughlin (1987), of Wright and Stone

By standardizing $f(\underset{\sim}{x})$, IRT based caution index ζ (Tatsuoka, 1985) will be

$$(13) \qquad \zeta = \frac{f(\underset{\sim}{x})}{\sum_{j=1}^{n} P_j(\theta_R)Q_j(\theta_R)\left[P_j(\theta_R) - T(\theta_R)\right]^2}$$

The mechanism of how the index ζ distinguishes atypical response patterns from typical ones is described in M. Tatsuoka and K. Tatsuoka (1986); so a more detailed explanation than that given in Section 7-2 will not be given in this paper. But the statistical properties of the bug distributions will be discussed later in the context of the theory changes mentioned earlier.

The next section introduces an elementary relational structure of the items that will be used for preparing a list of erroneous rules in the bug library.

5. Differential Ordering of Items by Underlying Cognitive Processes.

If a content domain for constructing a cognitive diagnostic test is determined, then logical task analysis can nominate cognitive attributes. The attributes may refer to production rules, procedural operations, item types, or, more generally, any cognitive subtasks. A set of n items can be characterized by K nominated attributes and expressed by K-element vectors. Let us call this matrix an attribute x item matrix, or Q-matrix (Embretsen, 1984). The initial task analyses could be carried out by several experts or master teachers. If two or more experts use different methods to solve a given set of problems, then they may get entirely different attribute x item matrices.

In this study, all the nodes of a directed process network will be called "attributes" and denoted by $A_1, A_2, \ldots A_K$. Items will be characterized by placing one in the (j,k) cell of the Q-matrix where j involves attribute A_k, and placing a zero when item j does not involve attribute A_k.

$$(14) \quad q_{kj} = \begin{cases} 1 \text{ if item j involves attribute } A_k. \\ 0 \text{ otherwise} \end{cases}$$

For example, in Figure 1b Method B is an attribute x item matrix for fraction subtraction problems solved by Method B (separate whole number and fraction parts).

Insert Figure 1 about here

The purpose of introducing an attribute x item matrix is, first, to make it easier to construct a set of items relevant to diagnosing misconceptions that result from a lack of knowledge or misunderstanding of an attribute or combination of several attributes. The second aim is to extract a set of binary patterns of n items from the matrix, each pattern being produced by a systematic application of an erroneous rule resulting from a misconception, incomplete knowledge of a targeted attribute, or a combination of several attributes.

Starting from such an attribute x item matrix, many researchers in a variety of disciplines such as biology, differential psychology, engineering, and sociology have investigated clustering techniques. Although the author has applied such techniques to the attribute x item matrix, the results were disappointing. By using a dataset,

Method A (Always convert mixed numbers to improper fractions.)

ATTRIBUTES	ITEMS																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1. Convert a whole number to a fraction	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	1	0
2. Convert 1st mixed number to fraction	0	0	0	1	1	0	1	0	1	1	1	0	1	1	0	1	1	1	1	1
3. Convert 2nd mixed number to fraction	0	0	0	1	1	0	0	0	0	1	1	0	1	1	0	1	0	1	0	1
4. Simplify before subtracting	1	0	0	1	1	0	0	0	0	1	1	1	0	0	0	0	0	1	1	1
5. Find a common denominator	1	1	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
6. Column borrow to subtract numerator	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	1	1	1
7. Reduce answer to simplest form	0	0	0	0	1	0	0	0	1	1	1	1	0	0	1	1	1	1	1	1

Method B (Separate mixed numbers into whole and fraction parts.)

ATTRIBUTES	ITEMS																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1. Convert a whole number to fraction or mixed number	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0
2. Separate whole number from fraction	0	0	0	1	1	0	1	0	1	1	1	0	1	1	1	1	1	1	1	1
3. Simplify before getting final number	0	0	0	1	1	0	0	0	0	1	1	1	0	0	0	0	0	1	1	1
4. Find the common denominator	1	1	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
5. Borrow 1 from whole number part, change numerator and whole	0	0	0	1	0	0	0	0	0	1	1	0	1	0	0	0	1	1	1	1
6. Column borrow to subtract 2nd numerator from 1st	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0
7. Reduce answer to simplest form	0	0	0	0	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0

Figure 1. Attribute x item matrices of fraction subtraction problems by using Methods A and B.

results of the analysis may be more objective, but interpretability of analysis results may be reduced to a great extent. The frequent lack of interpretability of factors obtained by factor analysis of the estimated trait variables is well known. Thus there is a problem of trade-off between establishing objectivity and interpretability. At this stage of making a diagnostic test, we will take the value of interpretability of data into account. However, it is not our intention to neglect objectivity. This issue will be discussed again in the summary and discussion.

Each item, represented by a column vector Q_j of the attribute x item matrix, is now characterized by a specific combination of attributes. Similarly, each attribute, a row vector $\underset{\sim}{Q}_k$ contains the information as to which items involve attribute k. Let R be a relation on a set of the column vector $\underset{\sim}{q}_j$, j=1,....,n, where n is the number of items.

The definition of R is as follows:

(15) $\underset{\sim}{q}_i \leq \underset{\sim}{q}_j$ if $q_{ik} \leq q_{jk}$ for k=1,....,K.

Or equivalently, if item j includes the attributes involved in item i, then item j requires more tasks than item i. This relation satisfies the reflexive and transitive laws:

1) $\underset{\sim}{q}_i \leq \underset{\sim}{q}_i$, (reflexive law);

2) if $\underset{\sim}{q}_i \leq \underset{\sim}{q}_j$ and $\underset{\sim}{q}_j \leq \underset{\sim}{q}_l$, then $\underset{\sim}{q}_i \leq \underset{\sim}{q}_l$ (transitive law).

With this relation, a set $\{\underset{\sim}{q}_1,....,\underset{\sim}{q}_n\}$ has a partial ordering. If the symmetric law is satisfied by $\underset{\sim}{q}_i$ and $\underset{\sim}{q}_j$, then $\underset{\sim}{q}_i$ and $\underset{\sim}{q}_j$ are

equivalent, and it is written as $\underset{\sim}{q}_i \sim \underset{\sim}{q}_j$. Let $\{\underset{\sim}{q}_j\}$ be a set of items expressed by vectors of K attribute elements where $\underset{\sim}{q}_j$, j=1,...n. From the attribute x item matrix a set of totally ordered items is extracted. When relation R exists for any two elements in a set S, then S is said to be totally ordered.

In Figure 1b, Method B, Items 6, 8, 12 and 10 are totally ordered but 6, 2, and 12 are not. Denote a set of totally ordered items by S_i. Then a list of the totally ordered sets of the items extracted from Figure 1, Method B, is presented in Table 1.

Insert Table 1 about here

In Figure 2b, Method B, a tree of the items is constructed from the list of totally ordered sets in Table 1. If two items are connected by a directed arc, then the items are totally ordered.

Insert Figure 2 about here

The number or numbers in a box represent the item(s) involving the attributes listed next to the box. An advantage to using such an item tree is that a structural interrelationship among the items can be portrayed schematically. With the process network it is difficult to see all the different solution paths taken by 20 items in a single graph, and it is also difficult to pinpoint why and where a student's response pattern deviates from the perfect responses. For example, Tatsuoka (1984) describes Rule 8 as "The student subtracts the smaller from the larger in corresponding parts when the two numbers are different." With Rule 8, items 12 and 14 (11/8 - 1/8,

Table 1
A List of Totally Ordered Sets of the Items Extracted from Figure 1, Method B

1.	6(8)*,2(3)*,1
2.	6(8) ,2(3) ,13
3.	6(8) ,12 ,5
4.	6(8) ,12 ,10
5.	6(8) ,14(16)*,17,13
6.	6(8) ,14(16) ,17,4(11,20)**,18,10
7.	6(8) ,14(16) ,17,4(11,20) ,19
8.	6(8) ,14(16) , 9,7,19

*Items 6 and 8, 2 and 3, 14 and 16 are equivalent.

**Items 4,11, and 20 are equivalent.

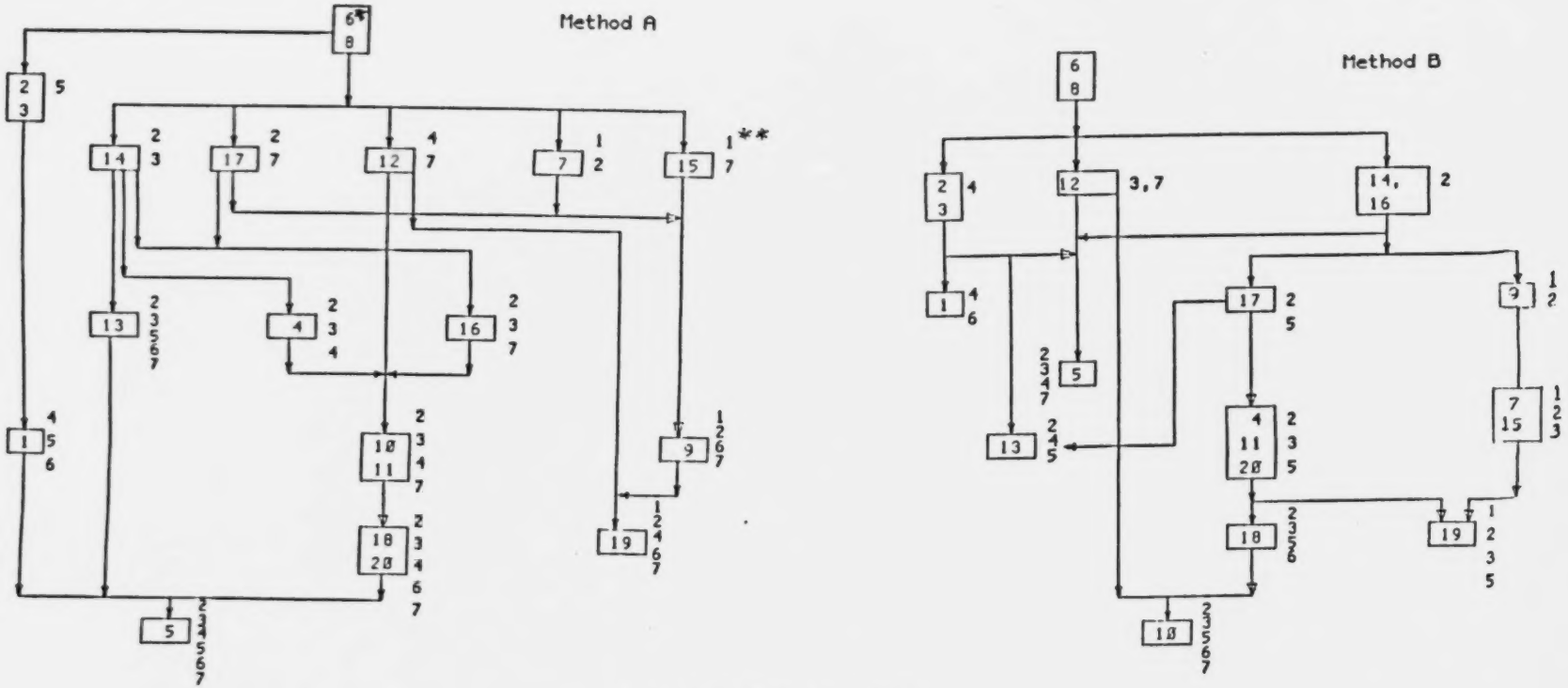


Figure 2. Two item trees constructed from the attribute x item matrices for Methods A and B.

3 4/5 - 3 2/5) have correct answers, while 2 and 17 (3/4 - 3/8, 7 3/5 - 4/5) don't. By marking each box in the item tree of Method B with * for the correct answers and x for the wrong ones, the sources of misconceptions producing Rule 8 (referred to as G14) are represented clearly, as can be seen in the following Figure 3; they are borrowing, getting a common denominator, and converting a whole number to a fraction.

Insert Figure 3 about here

6. Preparation of the "bug information bank" or "bug library."

The rule space model has been developed by emphasizing the importance of interpretability of statistics estimated from the data. Most psychological models have been less concerned with interpretability. Fischer and his associates (Embretson, 1985; Fischer, 1973), for example, expressed item difficulties in the Rasch model by a linear combination of component subtasks, and also unobservable frequencies with which each component influences the solution of each item. Since the models contain several parameters in the logistic functions, the estimation of the item parameters has become a major task in the past ten years (Fischer and Formann, 1972). Scheiblechner (1972) estimated item difficulties from the Rasch model first and then regressed the estimated item difficulties onto the hypothetical frequencies contained in the task matrix Q. Estimated β-weights approximate fairly well the estimates of component subtasks obtained by a conditional maximum likelihood procedure. However, Spada and McGaw (1985) state, "Despite the value

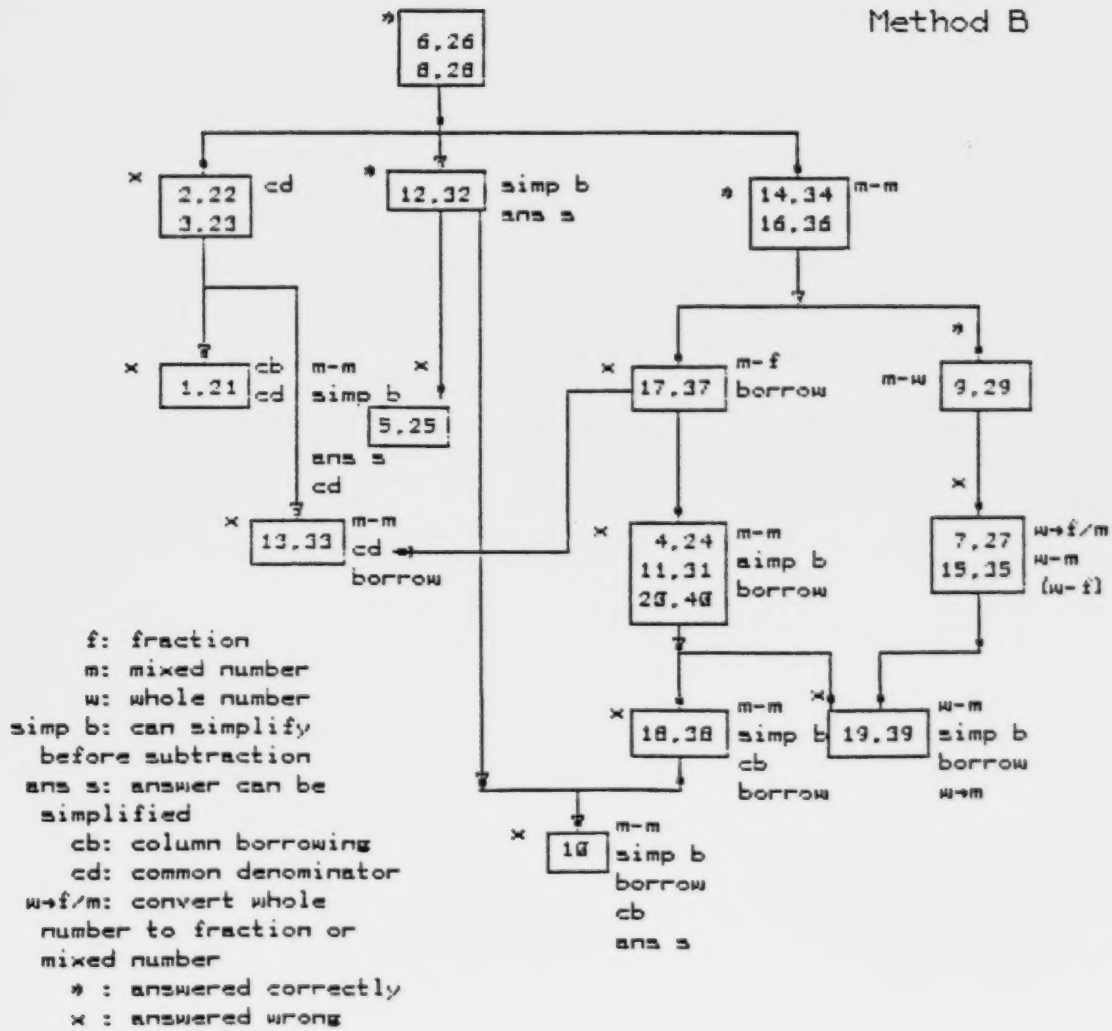


Figure 3. Representation of rule G14 by the item tree.

BEST COPY AVAILABLE

of the LLTM's analysis of task performance in terms of performance on elementary operations, there are some difficulties in interpreting the parameters of the model. The decomposition of the item difficulties is, of course, quite precisely defined but its psychological interpretation is not equally clear," (p. 180). Such difficulties in interpreting the estimates of the psychological models are very common in psychometrics. It is difficult to maintain the interpretability of estimated parameters in terms of underlying cognitive tasks in the current psychological modeling approaches. New approaches must be flexible enough so that individual differences unaccounted for during the process of formulating a model will not only improve the interpretability of the estimates but also be able to determine the existence of subjects who don't fit the model. Since the item trees are constructed from the inclusion relationships among attributes, interpretability is clearly retained.

6-1. Use of Multiple Regression

As did Scheiblechner (1972), a multiple regression analysis of the attributes on the item difficulties of 40 items was carried out. It showed that the four attributes--converting a whole number to a fraction or mixed number, getting the common denominator, borrowing of whole-number subtraction, and borrowing one from the whole number part to make the numerator larger--have significant β-weights to predict the item difficulties (Chevalaz, 1983). Also, the number of attributes involved in each item correlates with the item difficulty, ($r = .57$). Therefore, it is true that the greater the number of attributes involved in the items, the more difficult the items are.

Suppose that a student who uses Method A cannot get the lowest common multiple of two denominators but can do the remaining attributes. Then the response pattern of the performance on the 40 items will be 0 for the items involving the attribute cd in the Method A item tree and 1 for those not involving the attribute cd. Because getting the common denominator has a substantial β-weight, this error can be a good indicator for determining a list of rules in a bug library.

6-2. Use of the Item Tree

There are 2^7 possible response patterns obtainable from the attribute x item matrix in each method. However, the item trees in Figure 2 enable us to select a smaller number of rules and bugs that are substantially important in designing and evaluating lessons.

The numbers shown near the directed arcs of Figure 2 are conditional probabilities, Prob $(X_i = 1 | X_{i-1} = X_{i-2} = \ldots X_1 = 1)$ where i-1, i-2, ...,1 are antecedent items of i, and X_i is the score of item i.

Since $\underset{\sim}{q}_i \geq \underset{\sim}{q}_{i-1}$ (i.e., item i includes all the attributes involving item i-1) a drastic decrease in the value of the conditional probability implies that a newly added attribute (or attributes) causes the change. For instance, for Method B, the new attribute added to item 17 is indeed a difficult subtask--borrowing. If a student can't do this new attribute and can do the other attributes perfectly well, then subsequent items not including the borrowing attribute can be answered correctly. Thus, a binary response pattern corresponding to the student's performance will be

zeros for the items that involve borrowing and ones for non-borrowing items. This conjecture with respect to borrowing will be confirmed by examining the arc between items 12 and 10.

Next, the conditional probability value between items 9 and 7 is .60, which is low enough to merit attention. The new attribute in the second box is "w to f or m"--converting whole numbers to fractions or mixed numbers. Therefore, the second response pattern resulting from this case is zeros for the items with whole numbers such as 3 - 1/2, and ones for the items not including whole numbers in their first position. By proceeding in this manner a set of response patterns that are logically interpretable will be obtained. Thus, representing the structural relationships among the items with respect to their underlying cognitive attributes facilitates error analysis. We have developed a computer program to make an item tree from an attribute x item matrix and extract a set of totally ordered items. Then, applying the method just mentioned, a list of 39 rule classes is prepared. They are coded as G1 through G39 because some rule classes contain few rules. For instance, G2 is the binary pattern of ones for items 6, 8, 26, and 28, and G13 is binary pattern of ones for easy non-borrowing items. The interpretation of G2 is that a student can subtract two numbers if no attribute in Figure 1 is involved. Then their centroids and variance-covariance matrices are stored in the bug library for later use.

7. Stochastic Behavior of the Rules: Inferences from the Bug Distribution

7-1. Which rules are More Consistently Applied?

The bug distribution of Rule R defined on the neighboring response patterns of R was introduced in the previous section. It was expressed by a compound binomial distribution with the slip probabilities $S_j(\theta_R)$, j=1,...,n. Let $S(\theta_R)$ be the mean of $S_j(\theta_R)$, j=1,...n. Since any rule can be used as R, θ instead of θ_R will be used. Walsh (1954) expanded the compound binomial in powers of $S_j(\theta) - S(\theta)$. Suppose U is a random variable of slip and s is the number of slips, then the probability of having s slips is given by (16).

$$(16)\quad \text{Prob}(U=s) = P_n(s) + 1/2\ nV_2C_2(s) + 1/3\ nV_3C_3(s) + (1/4\ nV_4 - 1/8\ n^2V_2^2)$$
$$C_4(s) + (1/5\ nV_2 - 1/6\ n^2V_2V_3)C_5(s) + \ldots,\ s = 0,1,2,\ldots,n,$$

where $P_n(s) = \binom{n}{s} S(\theta)^s\ (1-s(\theta))^{n-s}$ for $s = 0,1,2,\ldots n,$

$$C_r(s) = \sum_{v=0}^{r} (-1)^{v+1} \begin{bmatrix} r \\ v \end{bmatrix} P_{n-r}(s-v),$$

$$V_r = 1/n \left[\sum_{g=1}^{r} \left[S_j(\theta) - S(\theta) \right] \right]^r,\ r = 2,3\ldots,n.$$

Let us use G2 and G13 for illustrating the relationship between the rule the probability of having s-slips away from the rules.

Insert Table 2 about here

As can be seen in Table 2, the probability of having a slip at

Table 2
Slip Probabilities of the First
20 Items for Rules G_2 and G_{13}

Item	G_2 Slip Probability	X_{R_j}	G_{13} Slip Probability	X_{R_j}
1	.043	0	.175	0
2	.025	0	.164	0
3	.025	0	.128	0
4	.154	0	.308	0
5	.285	0	.435	0
6	.610	1	.240	1
7	.006	0	.037	0
8	.495	1	.331	1
9	.430	0	.448	1
10	.002	0	.020	0
11	.012	0	.076	0
12	.269	0	.592	0
13	.001	0	.005	0
14	.192	0	.444	1
15	.006	0	.048	0
16	.208	0	.490	1
17	.003	0	.029	0
18	.021	0	.099	0
19	.000	0	.001	0
20	.001	0	.014	0

item 6 is .610 for Rule G2 and .240 for G13. Since the scores for item 6 of both G2 and G13 are 1, the probability of having a slip, 1 to 0, at item 6 is higher for G2 than for G13. That is,

Prob $(X_6 \neq X_{G2_6} \mid G2) = .610$ and Prob $(X_6 \neq X_{G13_6} \mid G13) = .240$.

Item 6 produces more slips for G2 than for G13. Table 3 shows the

Insert Table 3 about here

theoretical frequency distributions of Rules G13 and G2. The probability of having 5 slips for G2 is .225 and for G13 is .150. The frequency distribution of G13 reaches the mode at s = 6 while that of G2 reaches the peak at s = 5. A close examination of the bug distributions indicates that as θ comes closer to the sample mean, the rule associated with such θ produces more slips. As θ becomes larger or smaller, such rules produce fewer slips. This implies that the students with near-average θ values test their hypotheses more actively than those with θ-values close to the two extremes. Thus, theory changes are observed most often among the students who have not quite mastered a topic yet. This conjecture, derived from theoretical bug distributions conforms well with the results of error analysis performed on fraction addition problems as well as on signed-number subtraction problems (Birenbaum & Tatsuoka, in press).

7-2. Which Rules are Atypical?

The IRT models assume the local independence of reponses to the items. Therefore, the likelihood of each rule can be computed by Equation (17). For each rule R_i,

$$(17) \qquad L_{R_i} = \prod_{j=1}^{n} P_j(\theta)^{R_{ij}}(1 - P_j(\theta))^{1-R_{ij}}.$$

Table 3
Bug Distributions of Rules G_{13} and G_2 (N = 1000)

No. of Slips	Frequencies of G_{13}	Frequencies of G_2
0	1	1
1	5	11
2	18	48
3	49	122
4	98	198
5	150	225
6	181	188
7	176	119
8	141	58
9	93	22
10	51	6
11	23	1
12	9	
13	9	

It is known that the likelihood correlates very highly with ζ (Harnisch & Tatsuoka, 1983; Birenbaum, 1985, 1986). The numerator of ζ is the function $f(\underset{\sim}{x})$, which is linear in $\underset{\sim}{x}$, as follows,

$$(18) \qquad f(\underset{\sim}{x}) = \underset{\sim}{P}(\theta)\left[\underset{\sim}{P}(\theta) - \underset{\sim}{T}(\theta)\right] - \underset{\sim}{x}\left[\underset{\sim}{P}(\theta) - \underset{\sim}{T}(\theta)\right]$$

Since IRT curves are sensitive to the underlying cognitive processes as mentioned earlier (Tatsuoka, 1987), the deviation of $P(\theta)$ (denoted by $P(\theta) - T(\theta)$) also reflects those cognitive processes. But the shapes of the function $P_j(\theta)-T(\theta)$ will be different from those of the original logistic function. Figure 4 shows three curves where the difficulty of item 27 is smaller than the average function $T(\theta)$, and that of item 28 is greater than $T(\theta)$. Let us take $T(\theta)$ as the horizontal axis and the value of $P_j(\theta)$ as the vertical axis and draw the deviations of item response curves. Then graphs of deviations $P_j(\theta)$ will be as in Figure 5.

Insert Figures 4 & 5 about here

Easier items are located in the upper half of the space while more difficult items are in the lower half. If a student with ability θ takes the score of 1 for easier items and 0 for harder items, then the value $f(\underset{\sim}{x})$ will be smaller. If he/she scores 0 for easier items and 1 for harder items, then the values of $f(\underset{\sim}{x})$ tend to become larger. The same relation will hold for Equation (17), the likelihood function.

For example, if a student has a wrong rule for the borrowing operation in fraction subtraction, then his/her response pattern

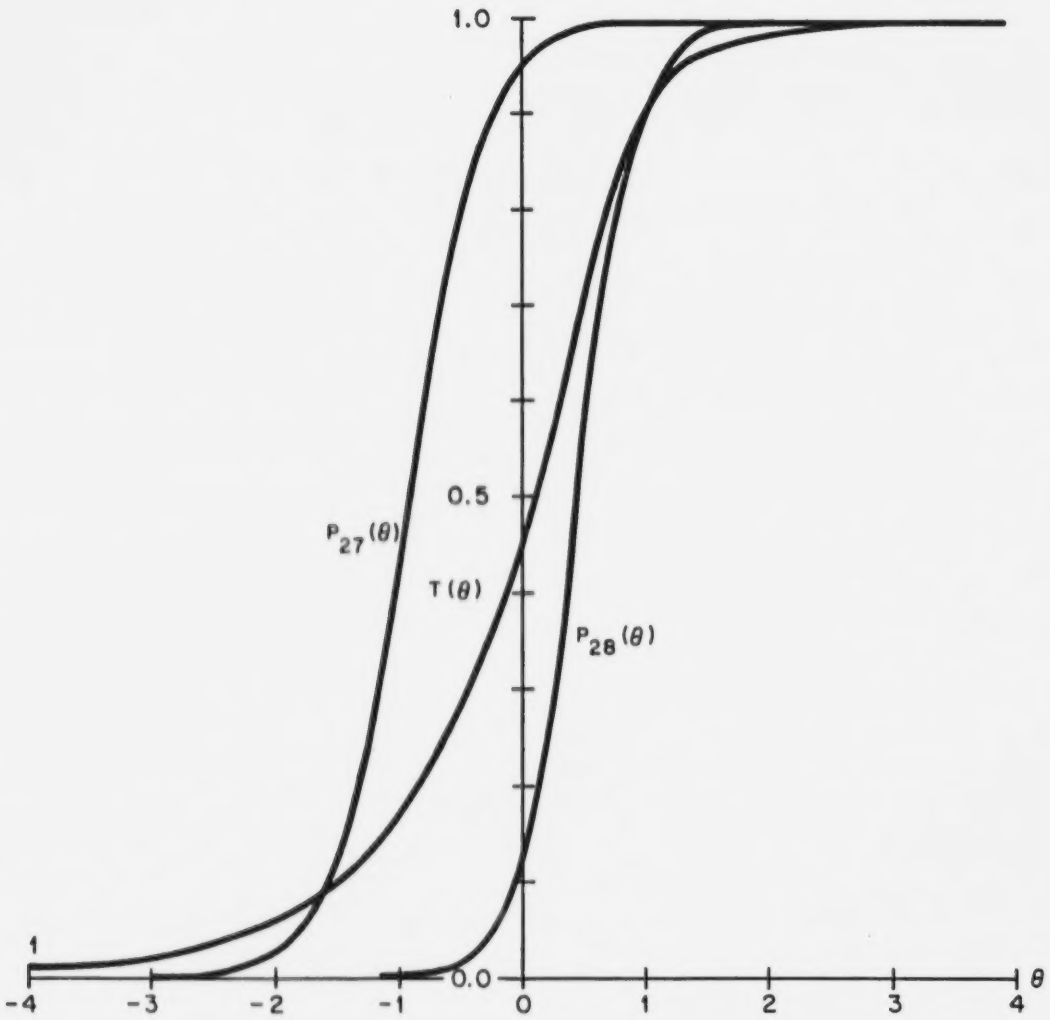


Figure 4. Average of 40 item response curves, T(Θ), and IRT curves of Items 27 and 28.

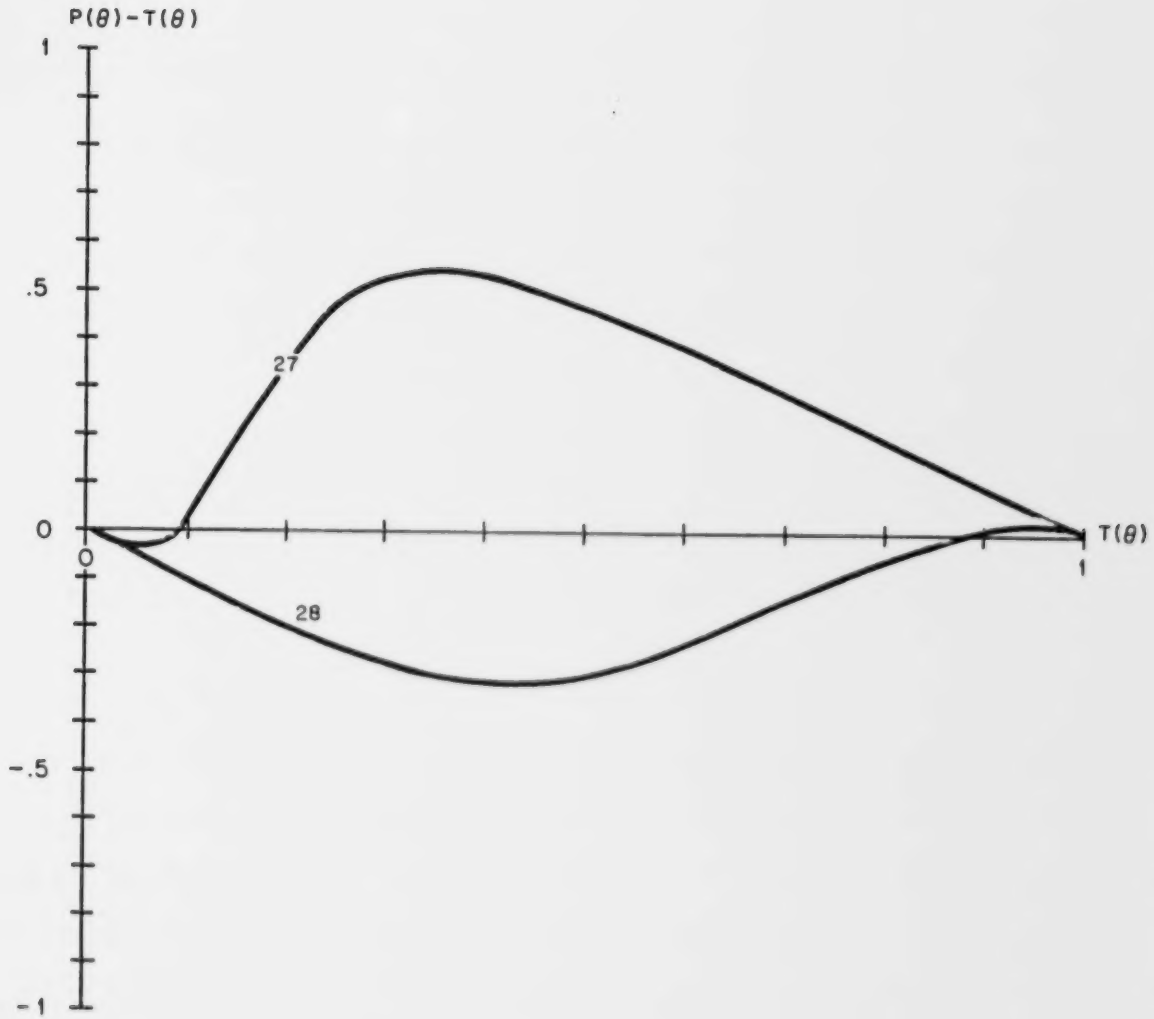


Figure 5. The curves of $P_{27}(0) - T(0)$ and $P_{28}(0) - T(0)$ over $T(0)$.

consists of 1's for the items which do not require borrowing and 0's for those requiring borrowing. Figures 6 and 7 show two sets of strikingly different curves of the deviation of $P_j(\theta)$. The first set of items, in Figure 6, contains items 4, 10, 11, 13, 17, 18, 19, and 20, which require borrowing before subtraction of the numerators is carried out. The second set of items, Figure 7, includes non-borrowing items 1, 2, 3, 5, 6, 8, 9, 12, 14, and 16; items 7 and 15 are excluded. Items 7 and 15 need conversion of a whole number to improper fraction or mixed number. All the items in Figure 6 (except item 4) have the functions $P_j(\theta)$-$T(\theta)$ below the x-axis which scales the average function $T(\theta)$. The functions in Figure 7 all have curves above the horizontal axis.

Insert Figures 6 & 7 about here

The response pattern associated with the borrowing error has the following binary vector:

G19 = (11101101100111010000111011011001110100 00)

and the maximum likelihood estimate of the latent variable θ is .09. The value of $T(\theta)$ for G19 is obtained by substituting θ = .09 into the item response functions and yields $T(.09) = .48$. Next, let us examine the values of the functions $P_j(\theta) - T(\theta)$ at $T(.09) = .48$. As can be seen in Figures 6 and 7, the functions $P_j(\theta)$-$T(\theta)$ at $T(\theta)$= .48 have clearly different values for borrowing than for non-borrowing items. Since the value of Equation (18) at a given θ depends on the item score of 1 or 0, the ζ value of G19 will be a negative number. By dividing the value of f(G19) by the standard

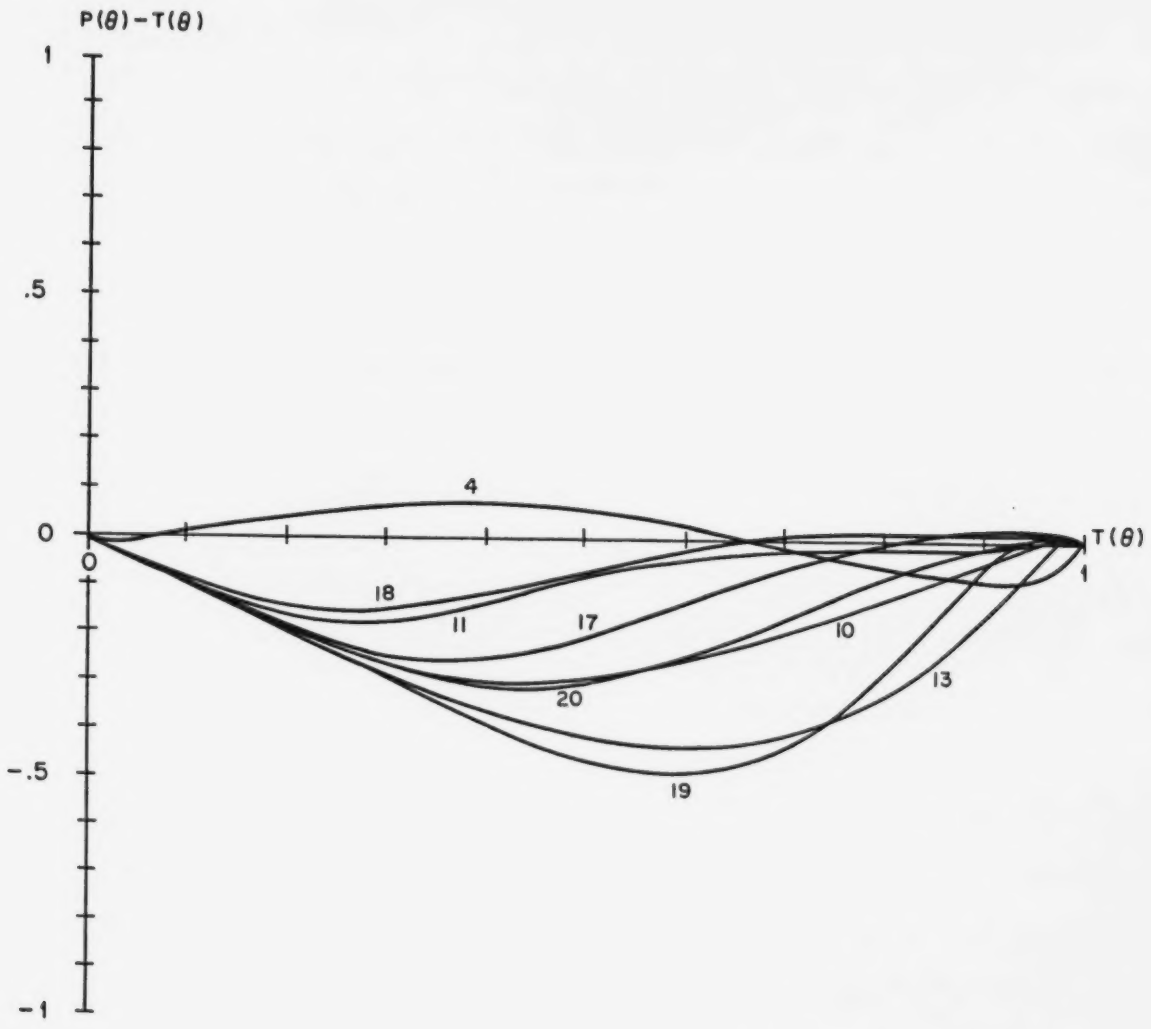


Figure 6. The curves of eight items requiring borrowing.

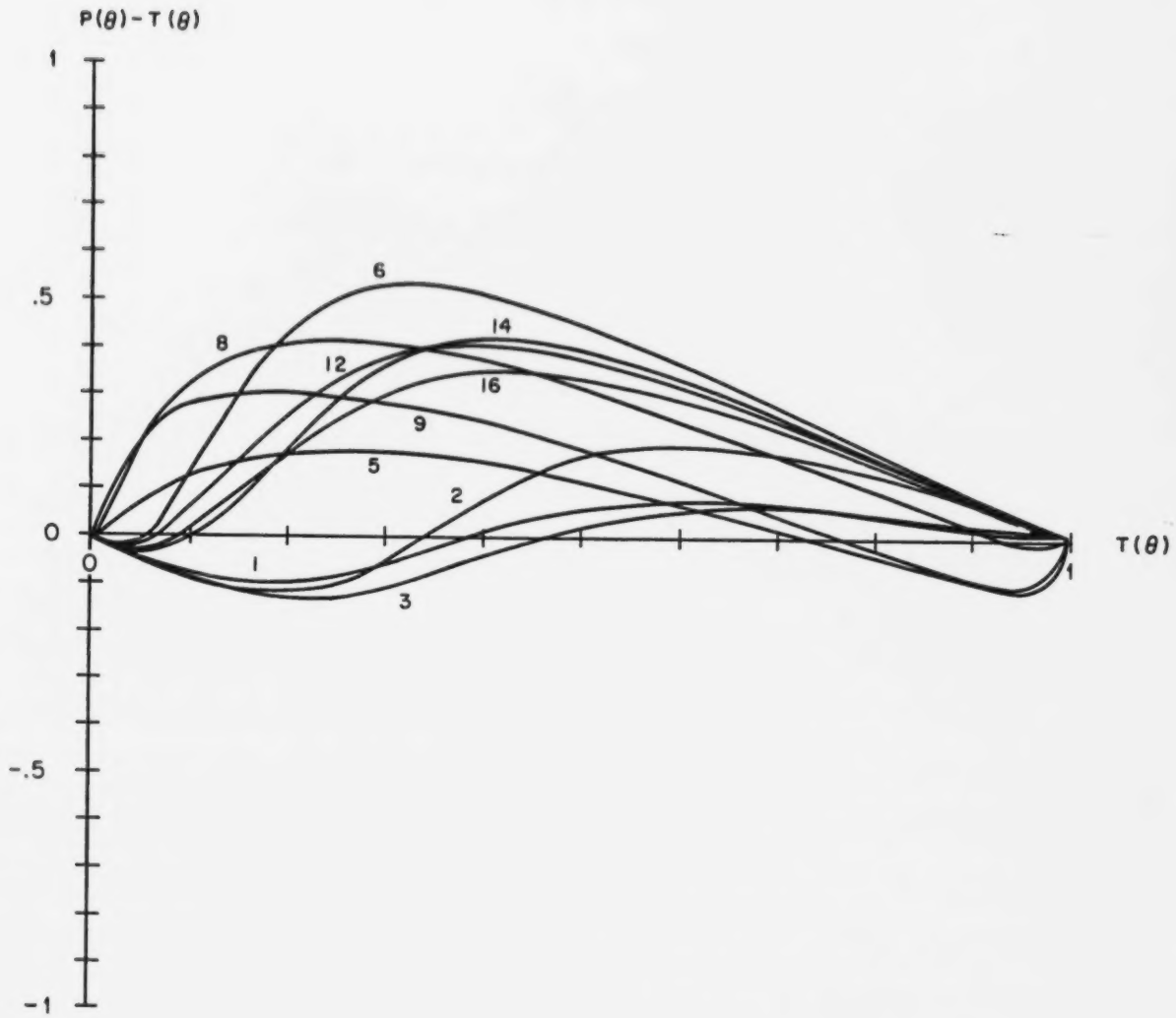


Figure 7. The curves of items NOT requiring borrowing.

deviation at θ, the value of Equation (13) is obtained. Thus, G19 corresponds to (.09, -2.26) in the rule space.

The rule space model starts by mapping all possible binary response patterns into a set of ordered pairs $\{(\theta,\zeta)\}$ and representing it by a Cartesian product space. A list of rule-class response patterns {Gr, r=1,...39} produced from the item-tree program (Baillie & K. Tatsuoka, 1985) will be expressed by a set of ordered pairs (θ_r, ζ_r). Table 4 lists the 39 points.

Insert Table 4 about here

Figure 8 shows the 39 centroids of 39 bugs in the rule space; their values are given in Table 4. In Figure 8, the cluster

Insert Figure 8 about here

represented by circled + signs are the rule classes or bugs derived from the item tree by Method A, and the squares are those by Method B. It is interesting to see that the two sets of bugs derived from the two entirely different structures of the item trees partition the rule space, and yet spread evenly over the θ-axis. Since the points below the θ-axis conform better to the order of item difficulties, they are more typical performances on the test items. If a rule is very unusual, then the location of the rule will be seen in an upper part of the space. So, the location within the space will tell whether or not the rule is atypical with high scores or low scores. The same is true for typical rules with high or low θ's. Figure 9 contains a selected set of ellipses whose major and minor axes are

Table 4
The 39 Centroids Representing 39 Different Error Types in Fraction Subtraction Tests (N = 535, n = 40)

Groups	θ	ζ	No. of Items	$I(\theta)^{-1}$	Rule Classes	θ	ζ	No. of Items	$I(\theta)^{-1}$
1	-2.69	-.80	1	.85	21	.24	-.89	22	.01
2	-1.22	-.69	4	.08	22	-.22	-1.23	14	.02
3	-.75	-.68	8	.05	23	.62	-1.55	32	.01
4	-.46	.75	10	.03	24	1.04	-.61	38	.03
5	.11	.91	18	.02	25	.75	-.05	34	.01
6	.64	1.74	30	.01	26	-.51	-1.62	10	.04
7	-.17	1.48	13	.02	27	-.87	-.56	6	.05
8	.40	-.16	25	.01	28	-1.99	1.01	2	.29
9	.60	-.43	31	.01	29	-.19	1.53	12	.02
10	.57	-.24	29	.01	30	-.24	2.74	10	.03
11	.99	.72	37	.03	31	-1.18	1.46	4	.07
12	1.19	.86	39	.05	32	-1.45	.58	4	.11
13	-.60	-1.58	10	.04	33	.57	-.66	31	.01
14	-.44	-2.31	12	.03	34	.59	-1.39	30	.01
15	-.18	.67	14	.02	35	-1.66	-1.96	4	.16
16	-.08	-1.81	16	.02	36	-.52	-.94	10	.04
17	.16	-.86	20	.02	37	-.32	-1.26	14	.03
18	-.01	-2.12	18	.02	38	-.41	-2.57	13	.03
19	.09	-2.26	20	.02	39	.17	-2.34	22	.01
20	.29	-1.51	24	.01					

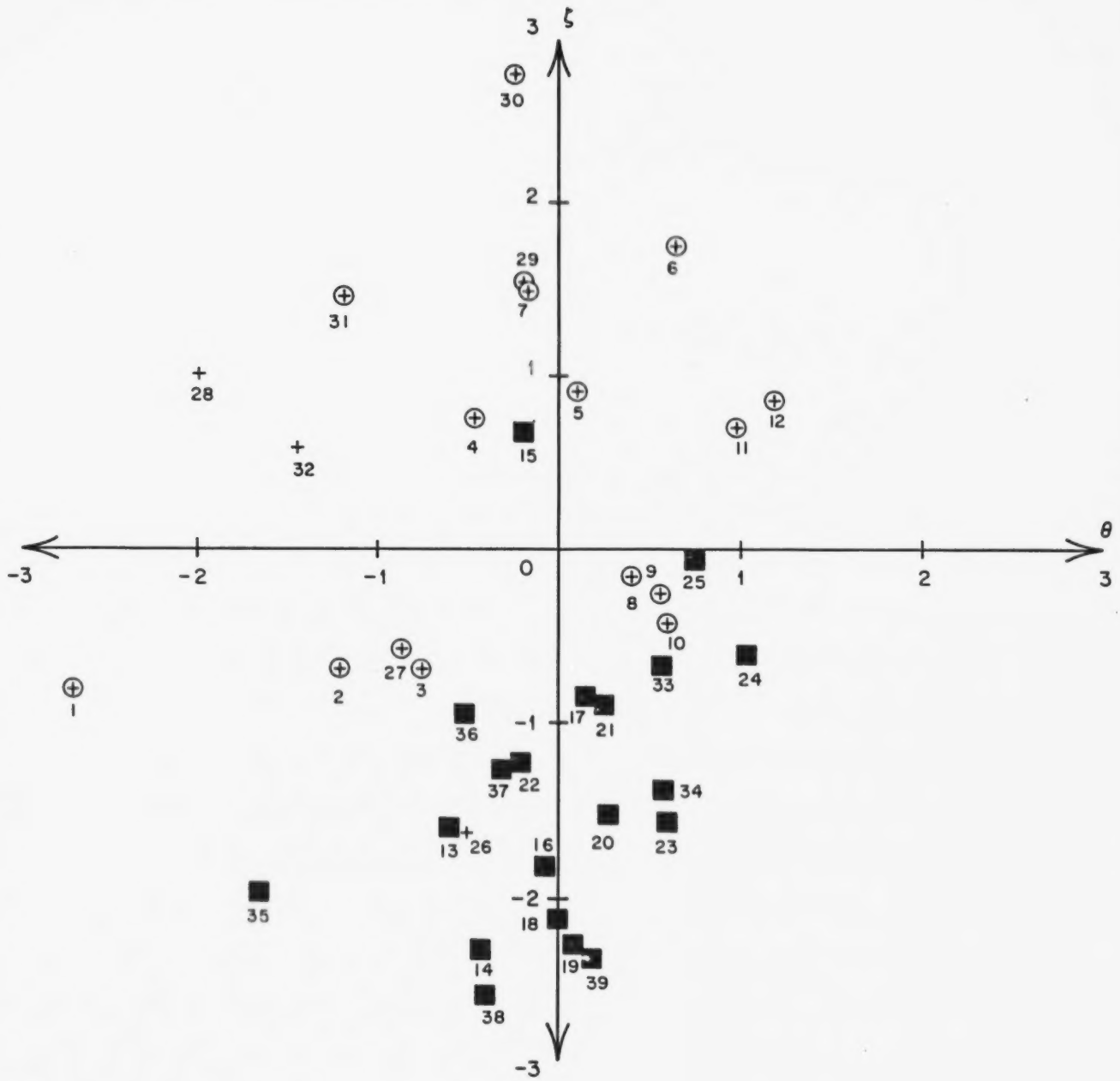


Figure 8. The centroids of 39 Bug Distributions. Note: Black squares are the bugs predicted from Method B. White circles are from Method A.

$1/I(\theta)$ and 1, respectively (Tatsuoka, 1985).

Insert Figure 9 about here

Operational Classification Scheme

If all rules prepared in the bug library are mapped into a set of ordered pairs, $\{(\theta_R, \zeta_R)\}$ along with their neighboring response patterns with several slips away from each rule, then the topography would be like Figure 9.

The population of points would exhibit modal densities at the rule points, and each rule forms the center of an enveloping ellipse with the density of points getting rarer as we depart farther from the center in any direction. Furthermore, the major and minor axes of these ellipses are parallel to the vertical and horizontal reference axes, respectively. The set of ellipses with Mahalanobis distance as the metric, gives a complete characterization of the rule space. If the ellipses represent misconceptions possessed by a majority of students, then any response-pattern point can be classified as most likely being a random slip from some rule. For a students' response-pattern point we search the two nearest ellipses by computing the Mahalanobis distances of the student's point to the centroids of the ellipses. Then, Bayes' decision rule for minimum error is applied to classify the point and to determine error probabilities.

However, computation of error probabilities (the probability of misclassifications) is not an easy task. If the two ellipses nearest the student's point have equal variance-covariance matrices and if

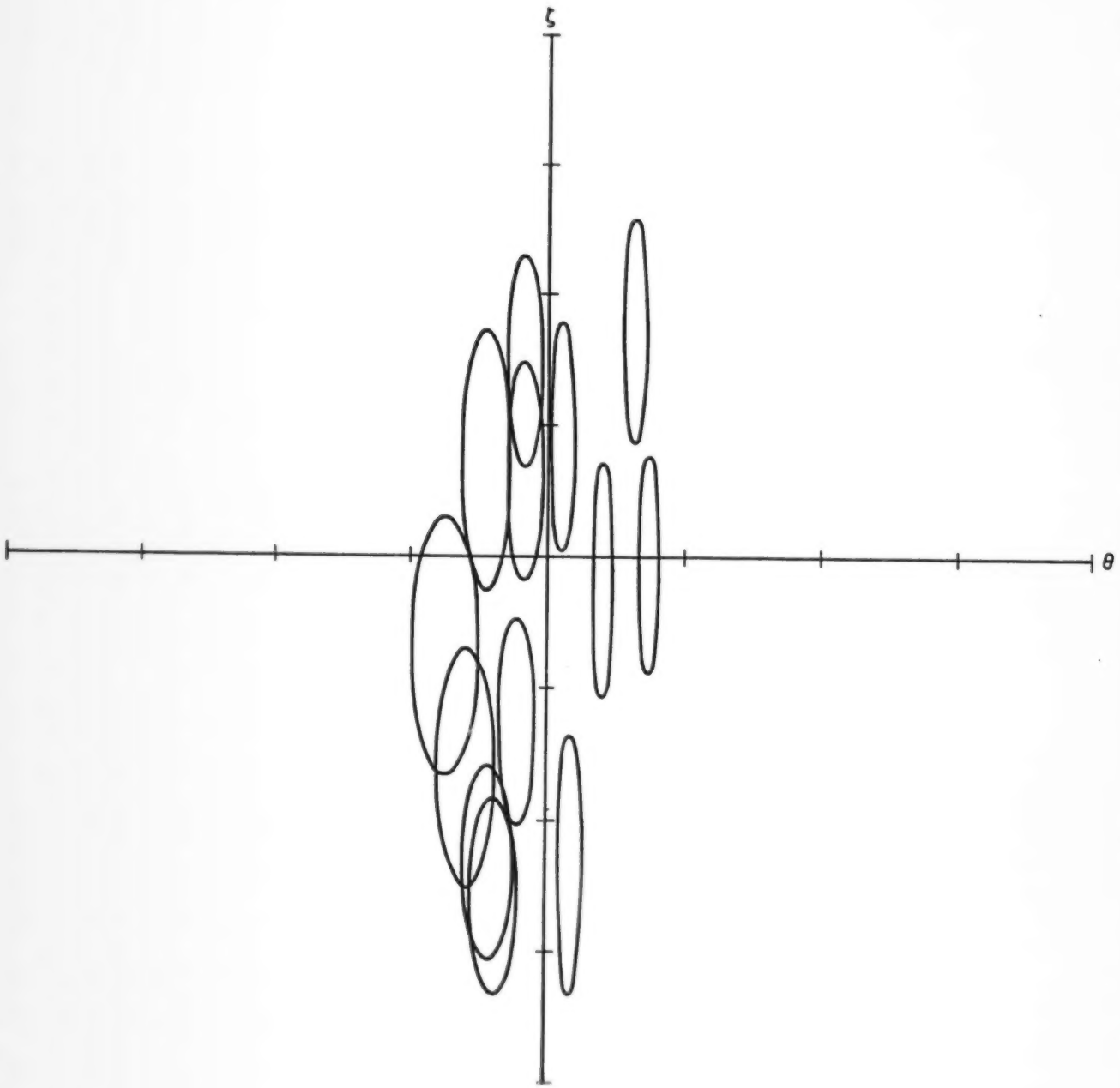


Figure 9. Fifteen ellipses representing fifteen error types randomly chosen from the 39 sets of ellipses.

the mapped distributions of the response patterns around the rules R_1 and R_2 into the rule space follow the multivariate normal distributions, then the logarithm of the likelihood-ratio function becomes linear. Therefore, computation of the error probabilities is reduced to the integration of the posterior conditional density functions (Tatsuoka & Tatsuoka, 1987a). More detail of the general procedure will be found in Fukunaga (1971). The assumption of equal variance-covariance matrices of R_1 and R_2 must, needless to say, be thoroughly examined.

1. The Results of Classification

The 535 students' responses on the 40-item fraction subtraction problems are mapped into the rule space, and almost 90% of the students are classified into 39 sources of misconceptions. Table 5 summarizes the number of students classified into the ellipses derived from the item trees of Method A and B in Figure 2.

Insert Table 5 about here

The distribution of Method A users and Method B users over the θ-axis, is fairly even and not much different in terms of the level of θ's. In an earlier study by Tatsuoka (1984), 275 out of 535 students, most of them seventh graders in a local junior high school, had never been taught Method B. The remaining 260 students were from a different school and most of them were in the eighth grade. Moreover, quite a few students used both methods. If they don't need borrowing, then they use Method B; otherwise they switch their method from B to A (Shaw, 1984, p. 43). Method A does not require the

Table 5
Frequencies of Students Who Used Either Method A or B

Range of θ	Method A	Frequencies	Method B	Frequencies
$\theta \leq -3$				
$-3 < \theta \leq -2.5$	1	29		
$-2.5 < \theta \leq -2$				
$-2 < \theta \leq -1.5$			36	12
$-1.5 < \theta \leq -1$	2,31	25		
$-1 < \theta \leq -0.5$	3,27	27	13,37	12
$-0.5 < \theta \leq 0$	4,7,29,30	42	14,15,16,18,22,38	39
$0 < \theta \leq 0.5$	5,8	46	17,19,20,21,39	31
$0.5 < \theta \leq 1$	6,9,10,11	78	23,25,34,35	73
$1 < \theta$	12	50	24	26
Total	16 rule classes	N=306	18 rule classes	N=193

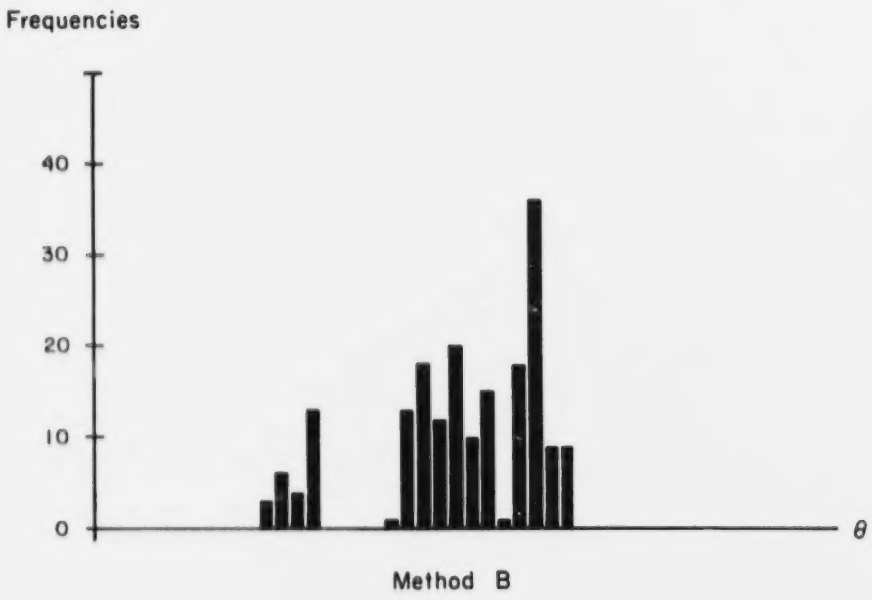
borrowing skill. These students are left undiagnosed in this study because the item trees utilized in producing a list of bugs are of either Method A or B and the combination tree of Methods A and B is not constructed. One of the advantages of using the rule space model is that uncertain errors on the test can be left classified without forcing them into one of the ellipses. They can subsequently be further investigated for their underlying cognitive processes. Therefore, the model can be exploratory and sensitive to any changes which affect students' performance.

Regarding the psychological meaning of θ, the content of Table 5 is investigated in detail. The frequency distributions of both groups over the θ-axis are shown in Figure 10.

Insert Figure 10 about here

The θ values in the two groups are not significantly different, but the ζ values are different. The two histograms for θ overlap and cannot be distinguished from each other. However, the Method-B users had a deeper understanding of the number system, and many of them were later advanced to an algebra class. If θ is really an ability influencing the scores of the 40-item fraction subtraction test, it is only natural to assume that the Method-B users should have higher ability levels than those of the Method-A users. As Resnick (1983) states, Method A requires a better short term memory with accurate computational skills, while Method B requires sophisticated manipulation of the numbers.

Tatsuoka, Linn, Tatsuoka and Yamamoto (1986) analyzed the



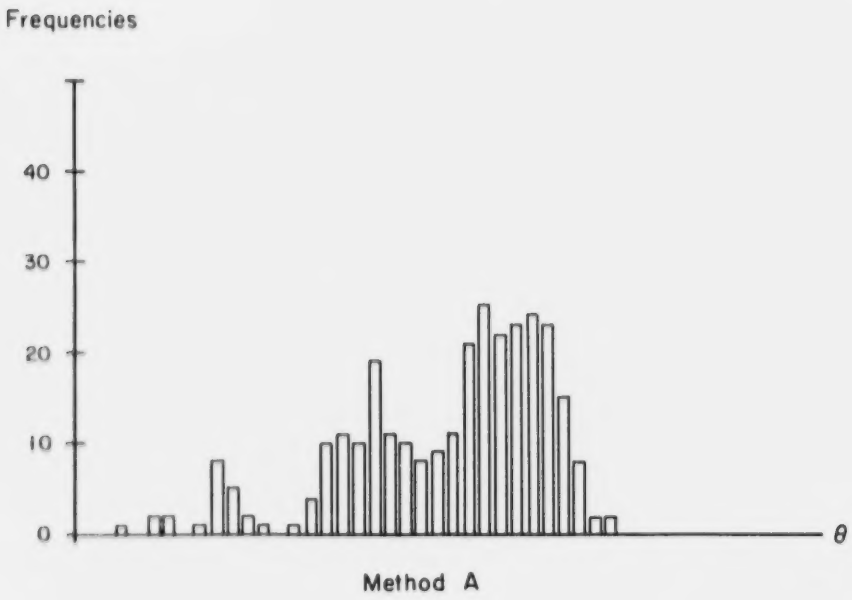


Figure 10. Frequency distributions for Method A and B users over Θ.

dataset more carefully and found that Method B users had higher mean scores for most subtasks except for borrowing, which clearly demonstrated that borrowing skills caused differential item performances. This study indicates that the rule space model is useful for studying item bias and for investigating the causes of item bias.

Summary and Discussion

This paper discussed some important issues for theories in educational measurement and testing. Recent findings in modern learning theory have raised tough challenges to psychometricians. Glaser categorizes these challenges into four main objectives which new achievement measurement should take into account. These objectives are to develop assessment methods that are descriptive, dynamic, structure-oriented, and that through automation make it possible to orchestrate several component tasks.

The pros and cons of two representative approaches of probabilistic modeling commonly used in psychometrics were discussed. Their common, basic principle is that an individual's proficiency level is explained, to a substantial degree, by defining certain human characteristics called traits. These traits are invisible; the only observable evidence of their existance is based on students' responses to the test items.

Modern physics and advances in the theory and practice of electricity and electronics share this problem with us. We have to infer the outcomes of invisible traits (if they exist), the unobservable cognitive processes, knowledge structures, and theory

changes from observable responses to test items. Modern physicists have discovered neutrons, electrons and other elementary particles by modeling observable physical phenomena as logical relationships.

It would seem that a new addition to the current theory and techniques of psychometrics is needed to respond to the new challenges raised by modern learning theory. This addition must overcome an ideological barrier as well as a technological one (Linn, 1985). At the same time, the new addition requires an expansion of our common sense to more abstract levels and a re-examination of the basic principles of test theories. The time is ripe for such developments, and finally research that manifests this new trend has begun to appear in several journals. This includes, for example, renewed concepts of construct validity proposed by Messick, Angoff, and Cronbach, and a new technology, called rule space, introduced by Tatsuoka and her associates.

This chapter introduced a part of the rule space model. The model consists of four major components: (1) item construction and preparation of the bug library, which consists of response patterns resulting from various sources of misconceptions, sources of incomplete knowledge, and erroneous rules of operation; (2) estimation of parameters, including item parameters of IRT models and bug distributions; (3) execution of decision rules; and (4) evaluation of the information in the bug library and update of the contents.

The first component was discussed mainly by introducing a method for constructing an item-tree. Making a tree is an application of

deterministic relational databases. Each item tree reflects a unique process structure underlying the problem-solving activities. The values of conditional probabilities computed on a specific directed path will help to identify the attribute that causes difficulties in doing the test.

After locating sources of error types, or combinations of misconceptions that require special attention in teaching, remediation, or designing instructions, the item-tree program converts them into a list of response patterns. Since the attributes can be production rules, the item tree can be a descriptive representation of a production system.

One of the main differences between the traditional modeling approach and the rule space approach lies in the ways they utilize the information obtained from a detailed task analysis. The former approach defines a set of new variables and formulates them as parameters in a probability function, or as probabilistic relationships among probability functions. The rule space approach utilizes algebraic relationships among item response functions for expressing the information obtained from the task analysis. Each rule corresponds to its unique bug distribution over slips and it is represented as true points in the rule space, which is a Cartesian product of two quantities, θ and ζ. Each rule is the center of an enveloping ellipse, with the density of points decreasing as we depart farther from the center in any direction. Further, the major and minor axis of these ellipses are asymptotically orthogonal (Tatsuoka, 1985). An observed response, on the other hand, will be

classified into one of the ellipses if possible. Statistical pattern recognition techniques are applied to classify the observed point. By examining the probability of errors, the student's performance on the test may be diagnosed in terms of a hypothetical knowledge structure and an interpretable prescription.

Since θ is taken as denoting the levels of proficiencies (and not as latent traits or constructs which determine the levels of test performance), the change scores resulting from hypothesis-testing activities of testees are explained (through a bug distribution of the random slip variable around rule R) without any philosophical difficulties. The rule space model has treated θ as a quantitative variable and attributed an important role to the contents of the bug library.

References

Alvord, G., & Macready, G. B.. (1985). Comparing fit of nonsubsuming probability models. Applied Psychological Measurement, 9(3), 233-240.

AMERICAN PSYCHOLOGICAL ASSOCIATION, AERA, and NCME. (1974). Standards for educational and pychological tests. Washington, DC: American Psychological Association.

Angoff, W. H. (1986). Validity: An evolving concept. A paper presented at the test validity conference, Educational Testing Service, Princeton, NJ.

Baillie, R. K., & Tatsuoka, K. K. (1985). Item Tree. A computer program. Urbana, IL: University of Illinois.

Birenbaum, M. (1985). Comparing the effectiveness of several IRT based appropriateness measures in detecting unusual response patterns, Educational and Psychological Measurement, 45, 523-534.

Birenbaum, M. (1986). Effect of dissimulation motivation and anxiety on response pattern appropriateness measure. Applied Psychological Measurement, 10(2), 167-174.

Birenbaum, M., & Tatsuoka, K. K. (1982). On the dimensionality of achievement test data. Journal of Educational Measurement, 19(4), 259-266.

Birenbaum, M., & Tatsuoka, K. K. (1986a). Effects of "on-line" test feedback on the seriousness of subsequent errors . (Research Report 86-3-ONR. Urbana, IL: University of Illinois, Computer-based Education Research Laboratory.

Birenbaum, M., & Tatsuoka, K. K. (in press). On the stability of students' rules of operation for solving arithmetic problems. Journal of Educational Measurement.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), Statistical theories of mental test scores, Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when the responses are scored in two or more nominal categories. Psychometrika, 37, 29-51.

Bock R. D., & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 46, 493-459.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. Applied Psychological Measurement, 6, 431-444.

Brown, J. S., & Burton, R. R. (1978). Diagnostic models for procedural bugs in basic mathematical skills. Cognitive Science, 2, 155-192.

Carroll, J. B. (1987). New perspectives in the analysis of abilities. Proceedings of the Buros-Nebraska Symposium on measurement and testing, the influence of cognitive psychology on testing and measurement. Lincoln, NE: University of Nebraska.

Cattell, R. B. (1971). Abilities: Their structure, growth, and action. New York: Houghton Mifflin.

Chevalez, G. M. (1983). A comparative analysis of two order analytic techniques: Assessing item hierachies in real and simulated data. Unpublished master's thesis, University of Illinois: Urbana, IL.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology revisited. American Psychologist, 30, 116-127.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal practical appropriateness indices. Applied Psychological Measurement, 11, 59-79.

Embretson, S.E. (1985). A general latent trait model for response processes. Psychometrika, 49, 175-186.

Fischer, G. H. (1973). The linear logistic test model as an instrument in education research. Acta Psychologica, 37, 359-374.

Fischer, G. H. (1981). On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. Psychometrika, 46, 59-77.

Fischer, G. H., & Formann, A. K. (1972). An algorithm and a FORTRAN program for estimating the item parameters of the linear logistic model (Research Bulletin No. 11). Vienna: Institute for Psychology, University of Vienna.

Fugkunaga, K. (1972). Introduction to statistical pattern recognition. New York: Academic Press.

Glaser, R. (1985). The integration of instruction and testing. Paper presented at the ETS Invitational Conference on the Redesign of Testing for the 21st Century, New York City.

Glaser, R., Lesgold, A., Lajoie, S. (1986). Toward a cognitive theory for the measurement of achievement (Technical Report)

Greeno, J. G. (1980). Instruction for skill and understanding in mathematical problem solving. Paper presented at the 22nd International Congress of Psychology, Leipzig.

Guilford, J. P. (1967). The nature of human intelligence. New York: McGraw-Hill.

Harnisch, D., & Tatsuoka, K. K. (1983). A comparison of appropriateness indices based on item response theory. In Hambleton (Ed.), Applications of Item Response Theory. Vancouver: ERIBC.

Klein, M., Birenbaum M., Standiford, S., & Tatsuoka, K. K. (1981). Logical error analysis and construction of tests to diagnose student "bugs" with addition and subtraction of fractions (Technical Report 81-6). Urbana, IL: University of Illinois, Computer-based Education Research Laboratory.

Lazarsfeld, P. F., & Henry, N. W. (1968). Latent structure analysis. Boston: Houghton-Mifflin.

Lee, T. T. (1983). An algebraic theory of relational databases. The Bell System Technical Journal, 62(10), 3159-3206.

Levine, M. V., & Drasgow, F. (1982). Appropriateness measurement: Review, critique and validating studies. British Journal of Mathematical and Statistical Psychology, 35, 42-46.

Linn, R. L. (1985). Barriers to new test designs. In E. E. Freeman (Ed.), Proceedings of the 1985 ETS Invitational Conference (pp. 69-79). The Redesign of Testing for the 21st Century: Princeton, NJ: Educational Testing Service.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. Psychological Reports, 3, 635-694. (Monograph Supplement 9).

Lord, F. M. (1953). An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. Psychometrika, 18, 57-77.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Macready, G. B. (1982). The use of latent class models for assessing prerequisite relations and transference among traits. Psychometrika, 47, 477-488.

Madaus, G. F., Woods, F., & Nutall, R. L. (1973). A causal model analysis of Bloom's taxonomy. American Educational Research Journal, 10, 253-262.

Marshall, S. P. (1980). Procedural networks and production systems in adaptive diagnosis. Instructional Science, 9, 129-143.

Mislevy, R. J. (1983). Item response models for grouped data. Journal of Educational Statistics, 8, 271-288.

Messick, S. (1980). Test validity and the ethics of assessment. American Psychologist, 35, 1012-1027.

Messick, S. (1984). The psychology of educational measurement. Journal of Educational Measurement, 21, 3, 215-238.

Miller, W. G., Snowman, J., & O'Hara, J. (1979). Application of alternative statistical techniques to examine the hierarchical ordering in Bloom's taxonomy. American Educational Research Journal, 16, 241-248.

Mokken, R. J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. Applied Psychological Measurement, 82, 4, 417-430.

Molemaar, I. (1987). The many null distributions of person fit indices (personal communication).

Mosier, C. I. (1941). Psychophysics and mental test theory, II. The constant process. Psychological Review, 48, 235-249.

Muthén, B. (1984). A general structural equation model with dichotomous ordered categories, and continuous latent variable indicators. Psychometrika, 49, 115-132.

Ohlsson, S., & Langley, P. (1985). Psychological evaluation of path hypotheses in cognitive diagnosis. In Mandl and Lesgold (Eds.), Learning issues for intelligent tutoring systems. New York: Springer.

Paulson, J. A. (1985). Latent class representation of systematic patterns in test responses (ONR Research Report). Portland, OR: Portland State University.

Ramsay, J. O. (1982). When the data are functions. Psychometrika, 47, 379-396.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. (Studies in mathematical psychology I.) Copenhagen: Neilsen and Lydiche (for Danmarks Paedogagiske Institut).

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. Applied Psychological Measurement, 9, 401-412.

Resnick, L. B. (1983). A development theory of number understanding. In Ginsburg (Ed.), The development of mathematical thinking. Orlando, FL: Academic Press, Inc.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika, Monograph Supplement, 34(4), part 2.

Samejima, F. (1983). A latent trait model for differential strategies in cognitive processes. Office of Naval Research Technical Report ONR/RR-83-1. Knoxville, TN: University of Tennessee.

Sceiblechner, H. (1972). Das lernen und lesen komplexer denkaufgaber. Zeitschrift fur Experimentelle und Angewandte Psychologie, 19, 476-506.

Shaw, D. (1984). Fraction subtraction errors: Case studies. In K. Tatsuoka (Ed.), Analysis of errors in fraction addition and subtraction problems, (NIE Final Report, pp. 40-51). Urbana, IL: University of Illinois, Computer-based Education Research Laboratory.

Shaw, D. J. (1986). The effects of using adaptive diagnostic test results as a basis for two types of computerized remediation. Unpublished doctoral dissertation. University of Illinois, Urbana, IL.

Shaw, D. J., Standiford, S. N., Klein, M., & Tatsuoka, K. K. (1982). Error analysis of fraction arithmetic -- selected case studies. (Research Report 82-2-NIE). Urbana, IL: University of Illinois, Computer-based Education Research Laboratory.

Sleeman, D. (1984). Basic algebra revisited: A study with 14-year-olds. (HPP-83-9, ED 258 846.) Stanford University, Department of computer science.

Snow, R. E. (1980). Aptitude and achievement. In W. B. Schrader (Ed.), New directions for testing and measurement: Measuring achievement: Progress over a decade. Proceedings of the 1979 ETS invitational conference. San Francisco: Jossey-Bass.

Snow, R. E., & Lohman, D. F. (1984). Toward theory of cognitive aptitude for learning from instruction. Journal of Educational Psychology, 76, 347-376.

Spada, H., & McGaw, B. (1985). The assessment of learning effects with linear test models. In S. E. Embretson (Ed.), Test Design: Developments in Psychology and Psychometrics (pp. 169-191). Orlando, FL: Academic Press, Inc.

Stout, W. (in press). A nonparametric approach for assessing latent trait unidimensionality. Psychometrika.

Sternberg, R. J. (1977). Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities. Hillsdale, NJ: Erlbaum.

Tatsuoka, K. K. (1975). Vector-geometric and Hilbert space reformulation of classical test theory. Unpublished doctoral disseration, University of Illinois, Urbana, IL.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. Journal of Educational Measurement, 20(4), 345-354.

Tatsuoka, K. K. (1984a). Analysis of errors in fraction addition and subtraction problems. (NIE Final Report). Urbana, IL: University of Illinois, Computer-based Education Research.

Tatsuoka, K. K. (1984b). Caution indices based on item response theory. Psychometrika, 49, 95-110.

Tatsuoka, K. K. (1984c). Changes in error types over learning stages. Journal of Educational Psychology, 76(1), 120-129.

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. Journal of Educational Statistics, 12(1), 55-73.

Tatsuoka, K. K. (1986). Diagnosing cognitive errors: Statistical pattern classification and recognition approach. Behaviormetrika, 19, 73-86.

Tatsuoka, K. K. (1987). Validation of cognitive sensitivity for item response curves. Journal of Educational Measurement, 24, 233-245.

Tatsuoka, K. K., Baillie, R., & Yamamoto, Y. (1982). SIGNBUG2: An error diagnostic computer program for signed-number arithmetic on the PLATO® system [Computer program]. Urbana, IL: University of Illinois, Computer-based Education Research Laboratory.

Tatsuoka, K. K., & Birenbaum, M. (1981). Effects of instructional backgrounds on test performances. Journal of Computer-based Instruction, 8, 1-8.

Tatsuoka, K. K., & Linn, R. L. (1983). Indices for detecting unusual patterns: Links between two general approaches and potential applications. Applied Psychological Measurement, 7(1), 81-96.

Tatsuoka, K. K., Linn, R. L., Tatsuoka, M. M., & Yamamoto, K. (1986). An application of the Mantel-Haenszel procedure to detect item bias resulting from the use of different instructional strategies. Manuscript submitted for publication.

Tatsuoka, K. K. & Tatsuoka, M. M. (1982). Detection of aberrent response patterns. Journal of Educational Statistics, 8(3), 215-231.

Tatsuoka, K. K., & Tatsuoka, M. M. (1983). Spotting erroneous rules of operation by the individual consistency index. Journal of Educational Measurement, 20(3), 221-230.

Tatsuoka, K. K., & Tatsuoka, M. M. (in press). Bug distribution and pattern classification. Psychometrika.

Tatsuoka, M. M., & Tatsuoka, K. K. (1986). Rule space. In Kotz and Johnson (Eds.)., Encyclopedia of Statistical Sciences. New York: Wiley.

Tatsuoka, K. K., & Tatsuoka, M. M. (1987b). Item response theory, latent classes and rule space. New technical report.

Trabin, T. E., & Weiss, D. J. (1979). The person response curve: Fit of individuals to item characteristic curve models. (Research Report 79-7-ONR). Minneapolis, MN: University of Minnesota, Department of Psychology.

VanLehn, K. (1983). Felicity conditions for human skill acquisition: Validity an AI-based theory (Research Report CIS-21-ONR). Palo Alto, CA: Cognitive & Instructional Sciences Group, XEROX Research Center.

Walsh, J. E. (1954). Approximate probability values for observed number of "successes" from statistically independent binomial events with unequal probabilities. Sankhya, 15, 281-290.

Wright, B. D., & Stone, M. H. (1979). Best Test Design, MESA Press, Chicago.

Yamamoto, Kentaro (1987). A model that combines IRT and latent class models. Unpublished doctoral dissertation. University of Illinois at Urbana-Champaign.